R 语言科研数据可视化

王敏杰

2020 年 9 月 25 日

四川师范大学

内容

- 基本图形
  - 常用图表类型
- 绘图原则
  - 绘图原则
  - 错题解析
- 代码实现
  - 常用工具
  - R 和 ggplot2
  - 案例演示

基本图形

基本图形

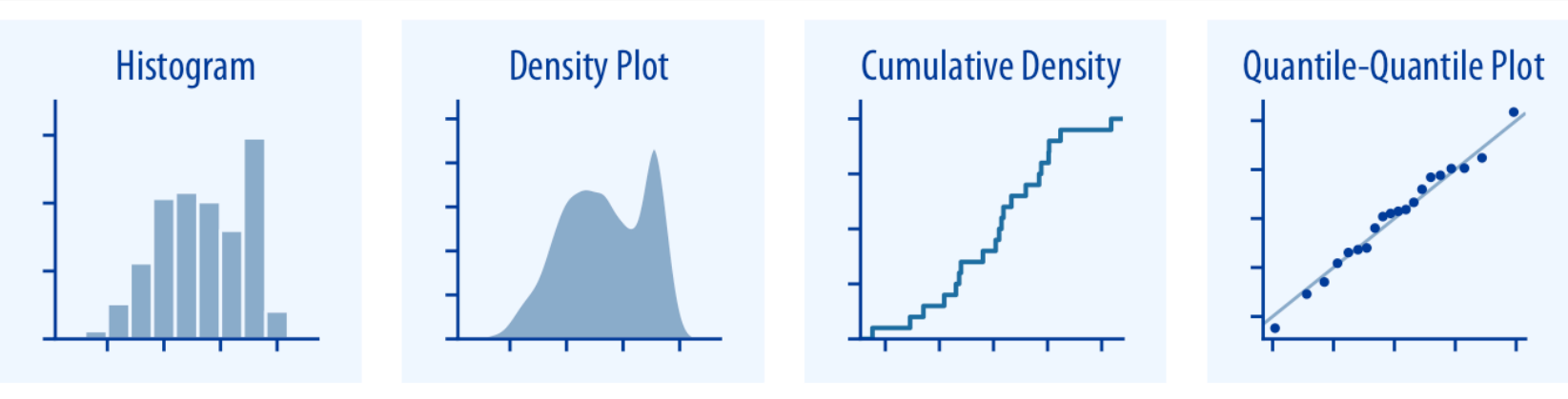
用什么样的图，取决于我们的需求

- 呈现数量
- 呈现分布
- 呈现比例
- 呈现关联
- 呈现不确定性
- 呈现时间变化
- 呈现地理信息

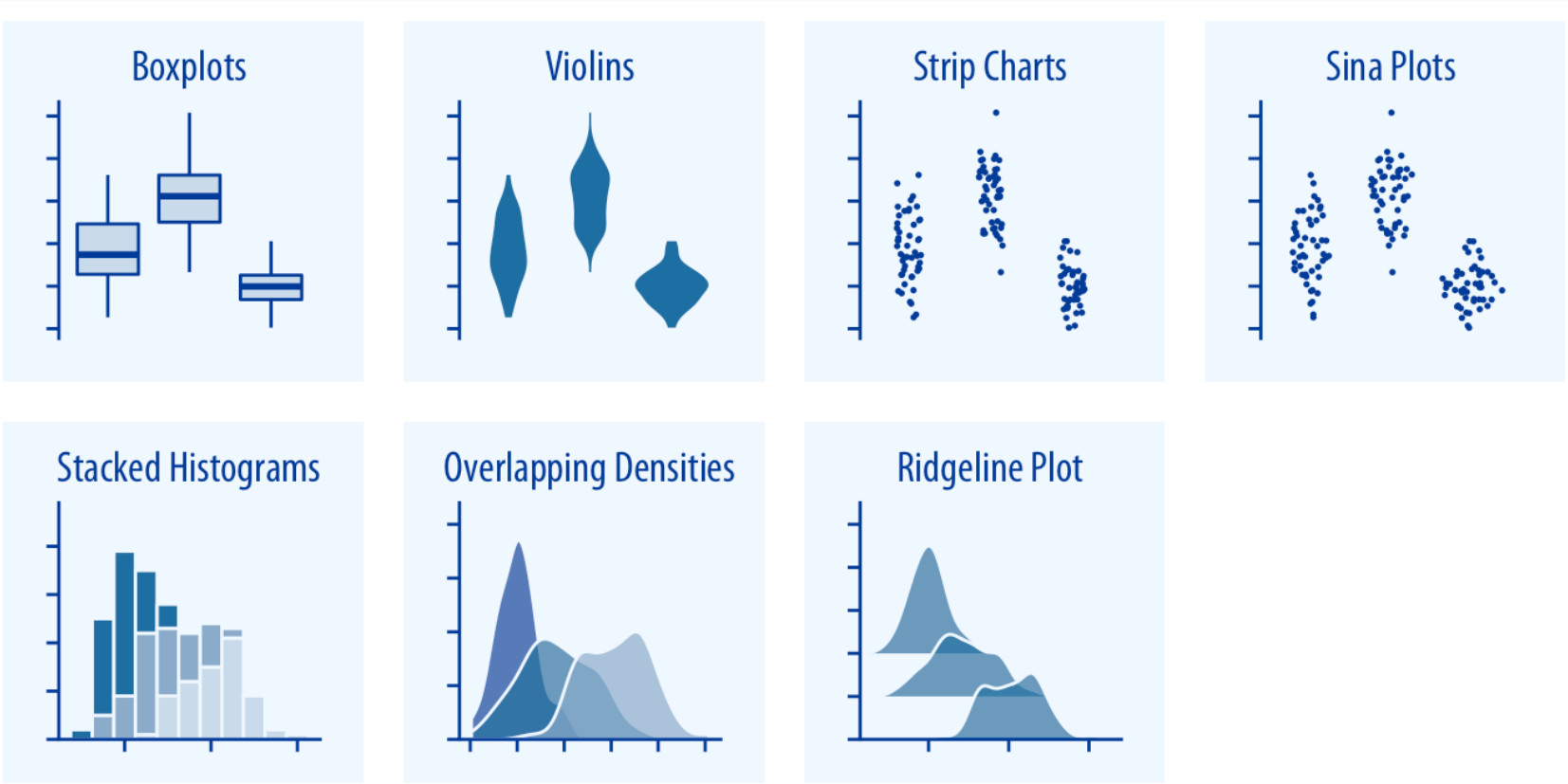
呈现数量


Bars
Bars
Dots
Grouped Bars
Grouped Bars
Stacked Bars
Stacked Bars
Heatmap

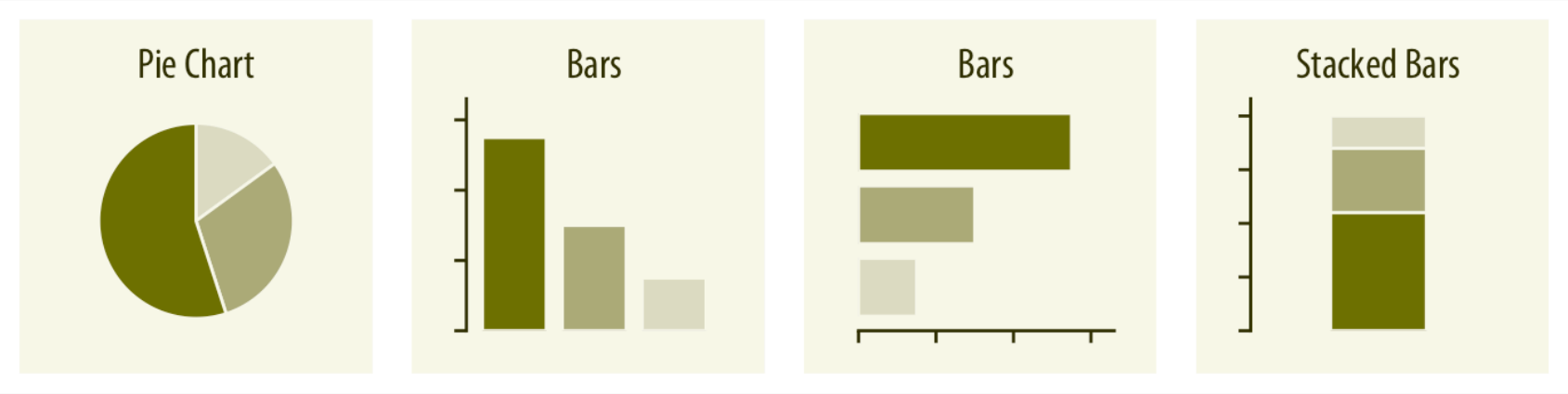
分布图


Histogram
Density Plot
Cumulative Density
Quantile-Quantile Plot

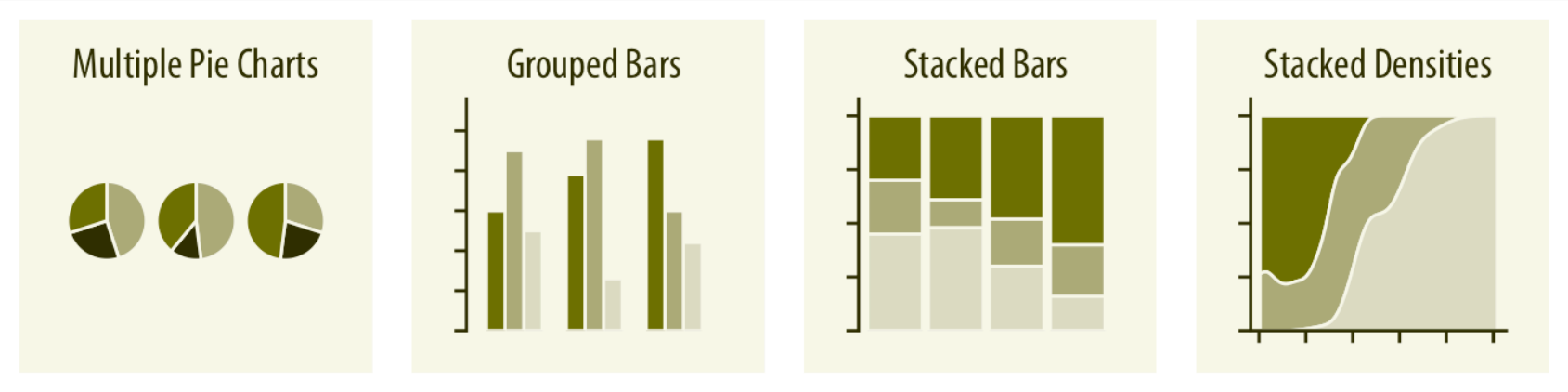
分布图


Boxplots
Violins
Strip Charts
Sina Plots
Stacked Histograms
Overlapping Densities
Ridgeline Plot

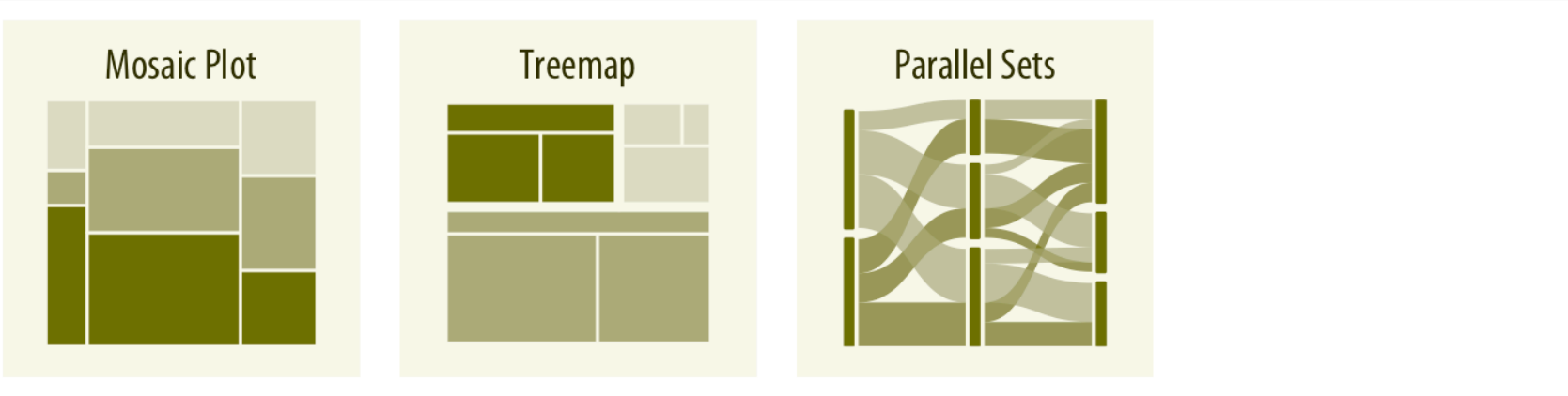
呈现比例



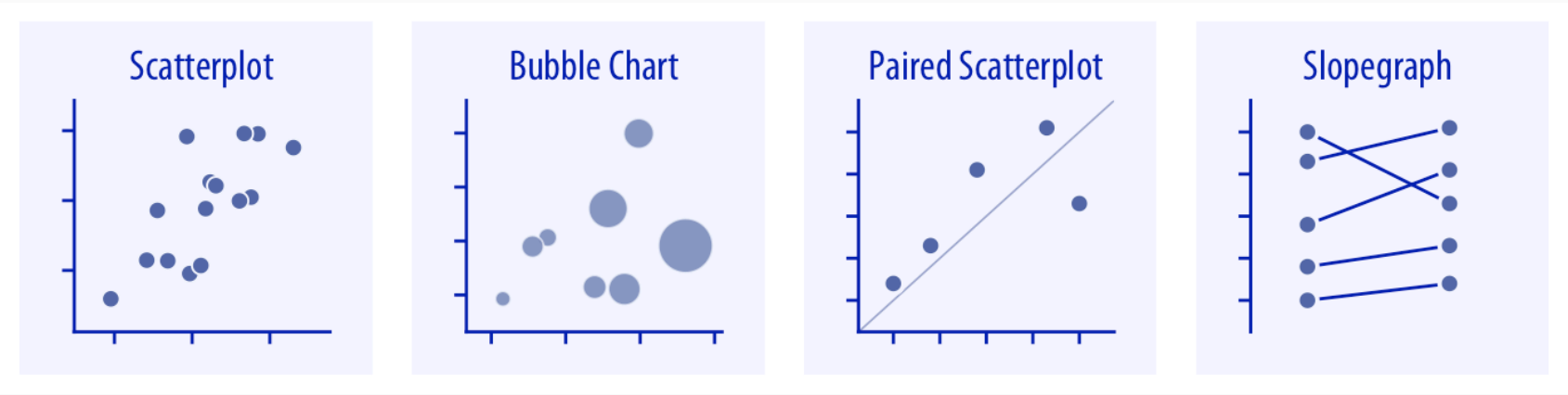
呈现比例


Multiple Pie Charts
Grouped Bars
Stacked Bars
Stacked Densities

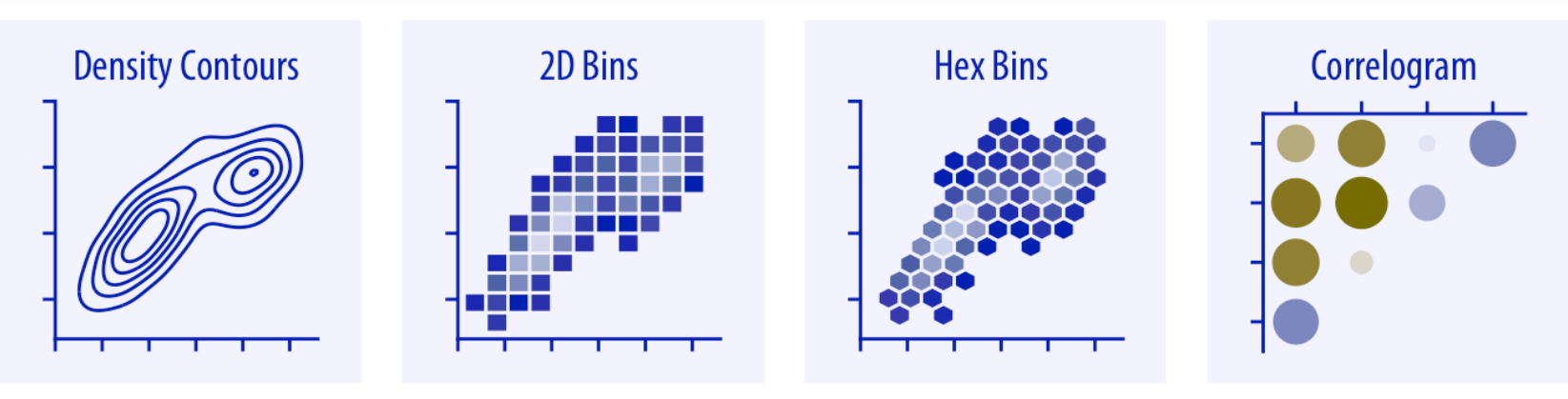
呈现比例



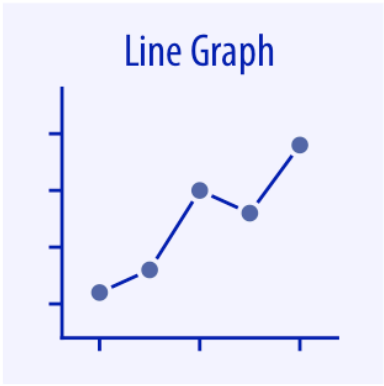
x–y 关联图

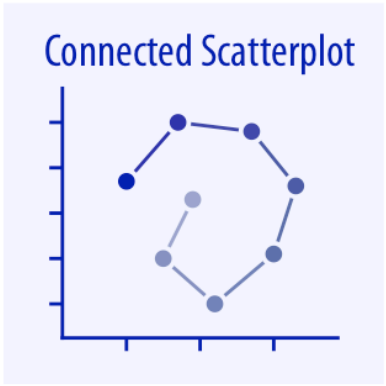


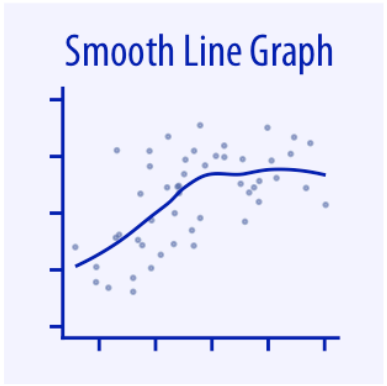
x–y 关联图



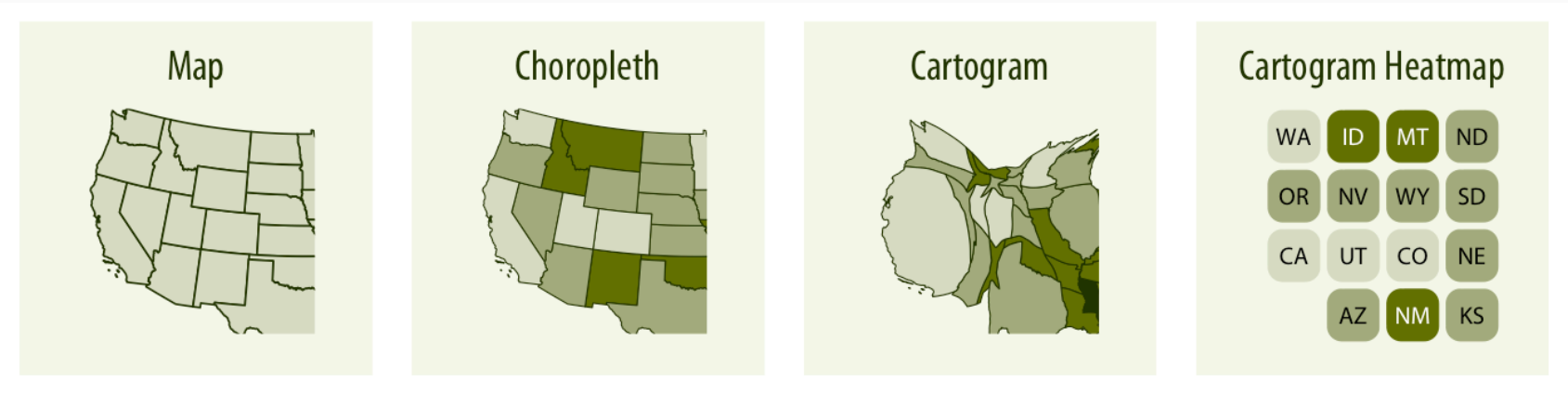
x–y 关联图



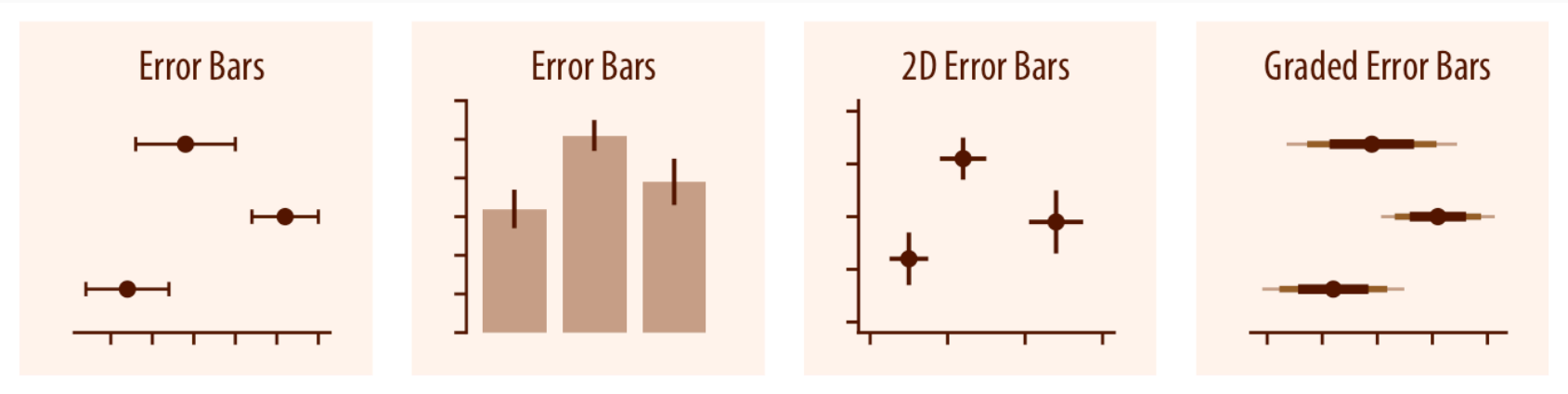




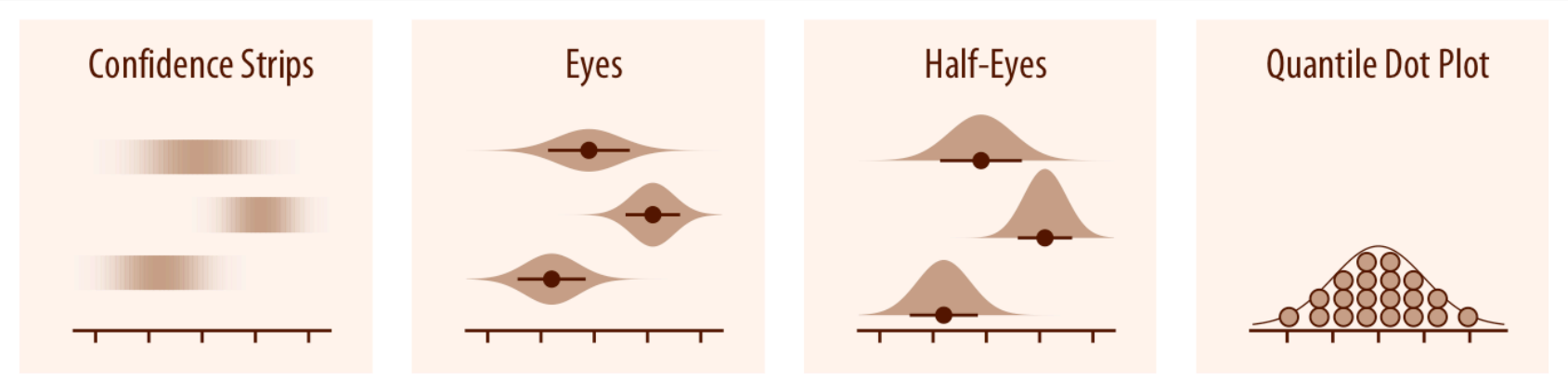
地图


Map
Choropleth
Cartogram
Cartogram Heatmap
WA ID MT ND
OR NV WY SD
CA UT CO NE
AZ NM KS

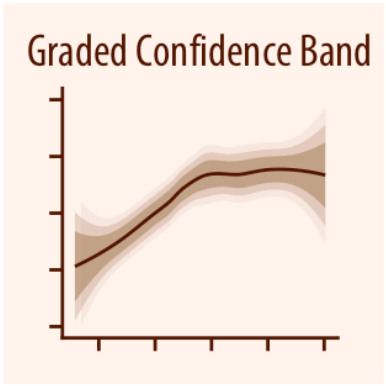
呈现不确定性

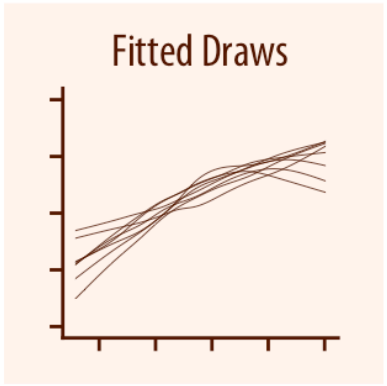


呈现不确定性



呈现不确定性



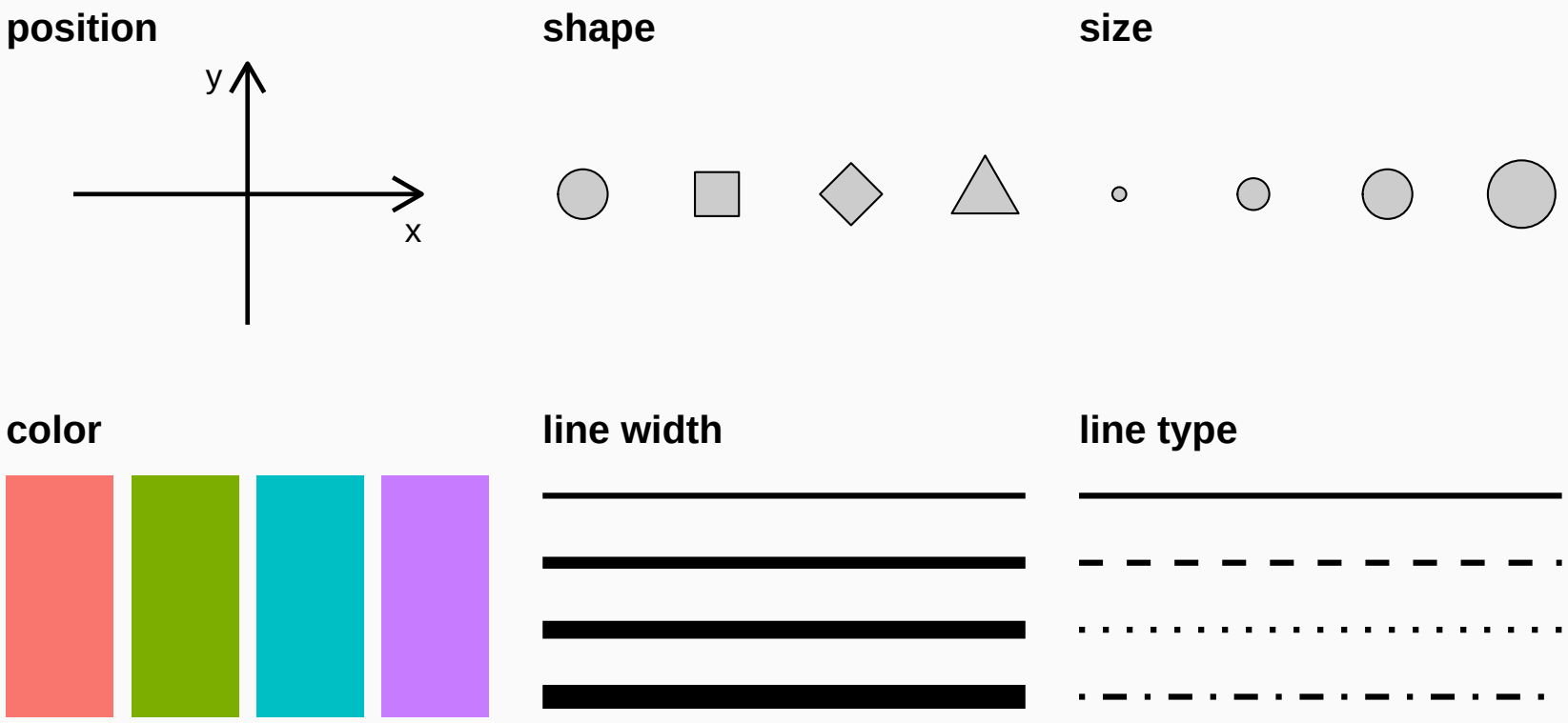


基本原则

基本原则

准确传递信息，同时不增加读者心智负担

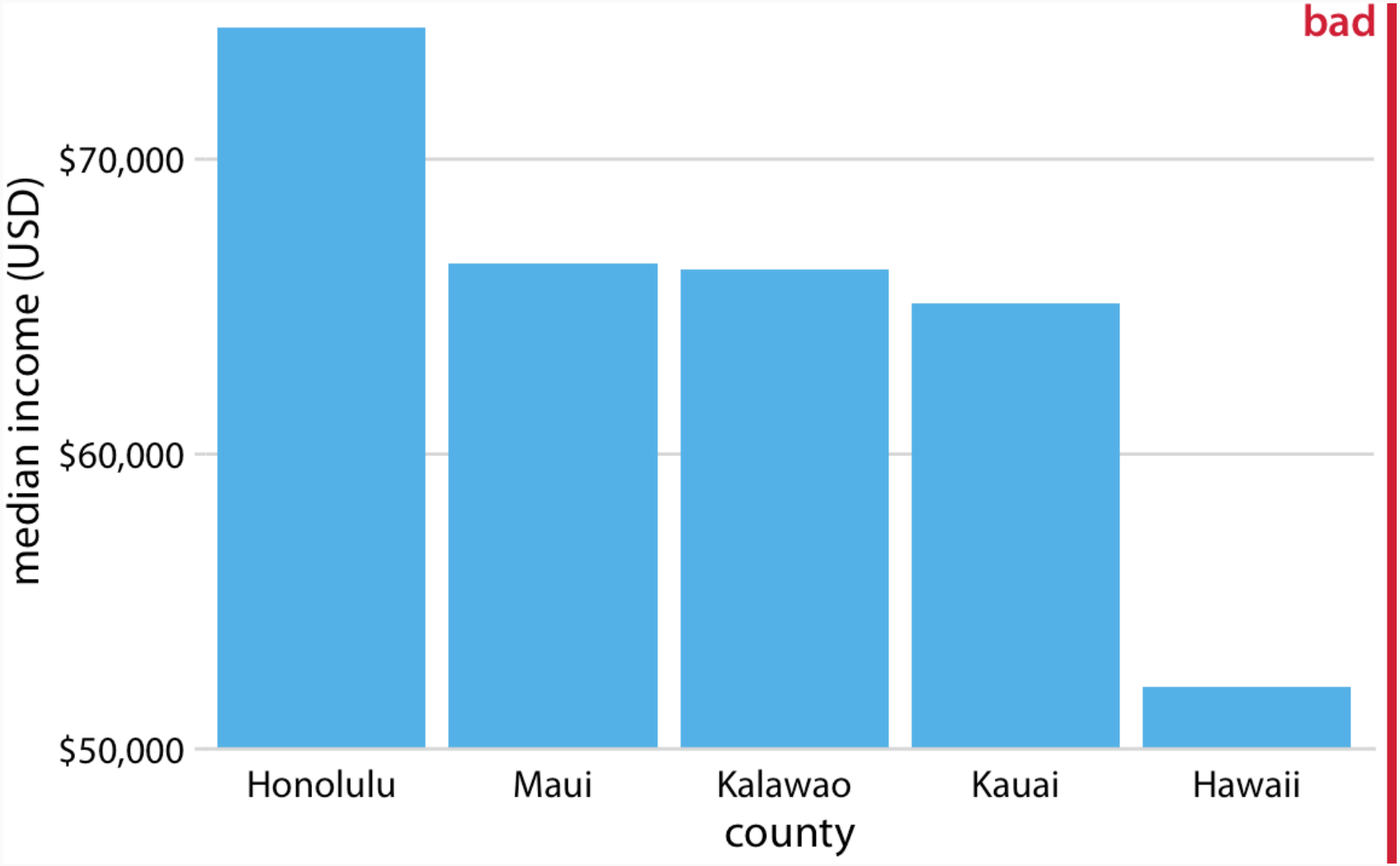
比例对等原则



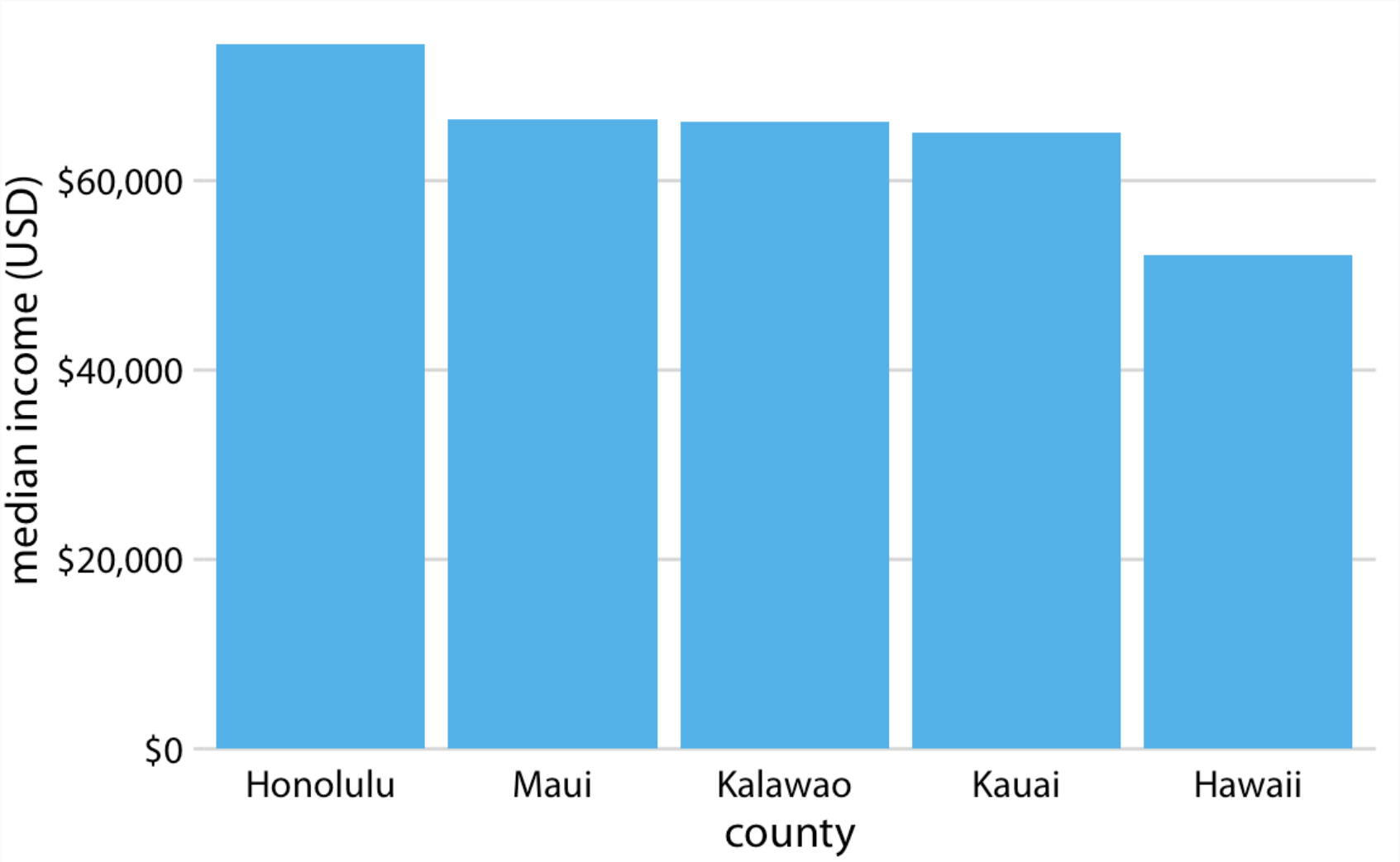
比例对等原则 (The principle of proportional ink)：

- 图形元素的面积越大，代表的数值越大，成一定比例。

比例对等原则

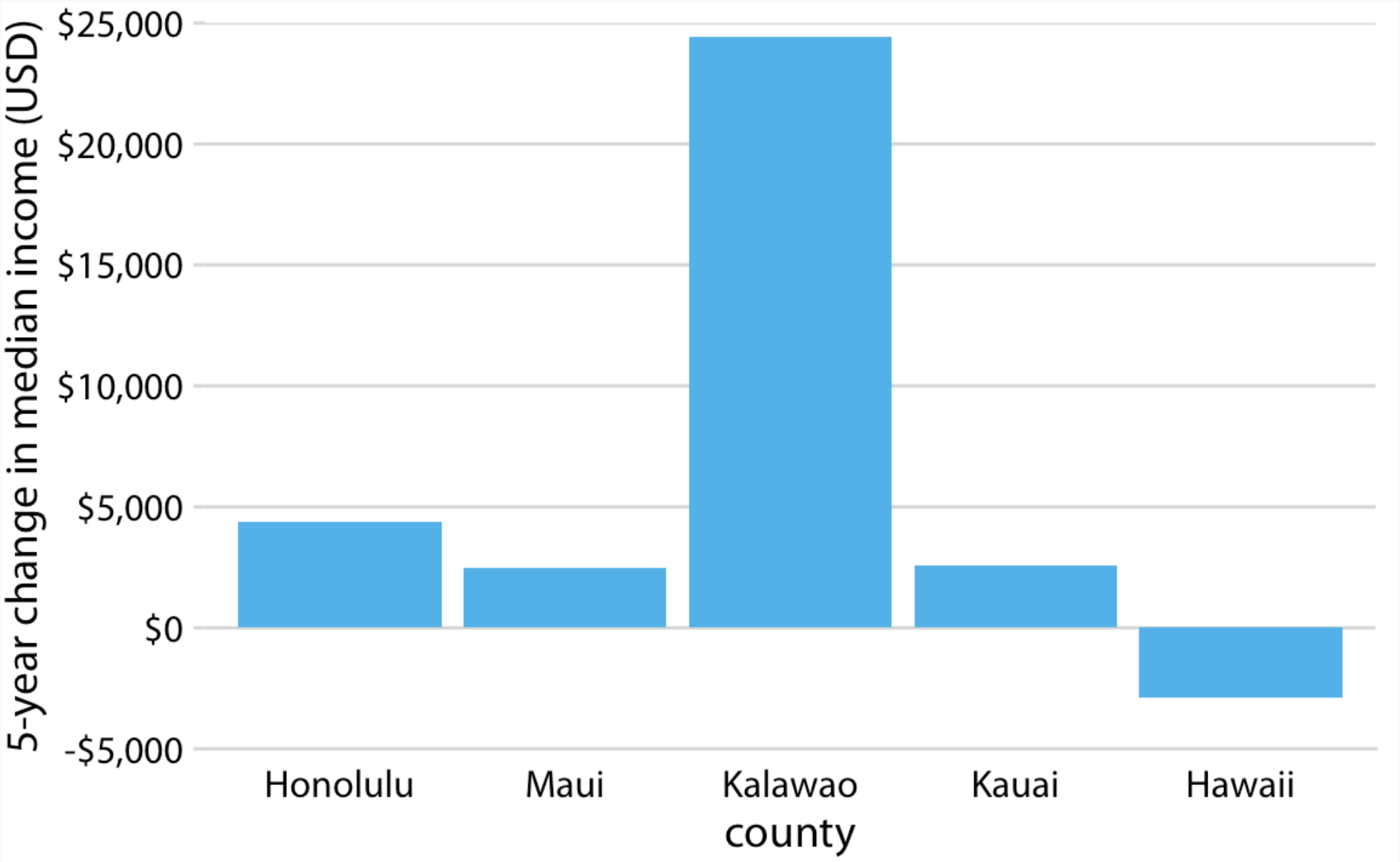

bad
median income (USD)
$70,000
$60,000
$50,000
Honolulu
Maui
Kalawao
Kauai
Hawaii
county

比例对等原则

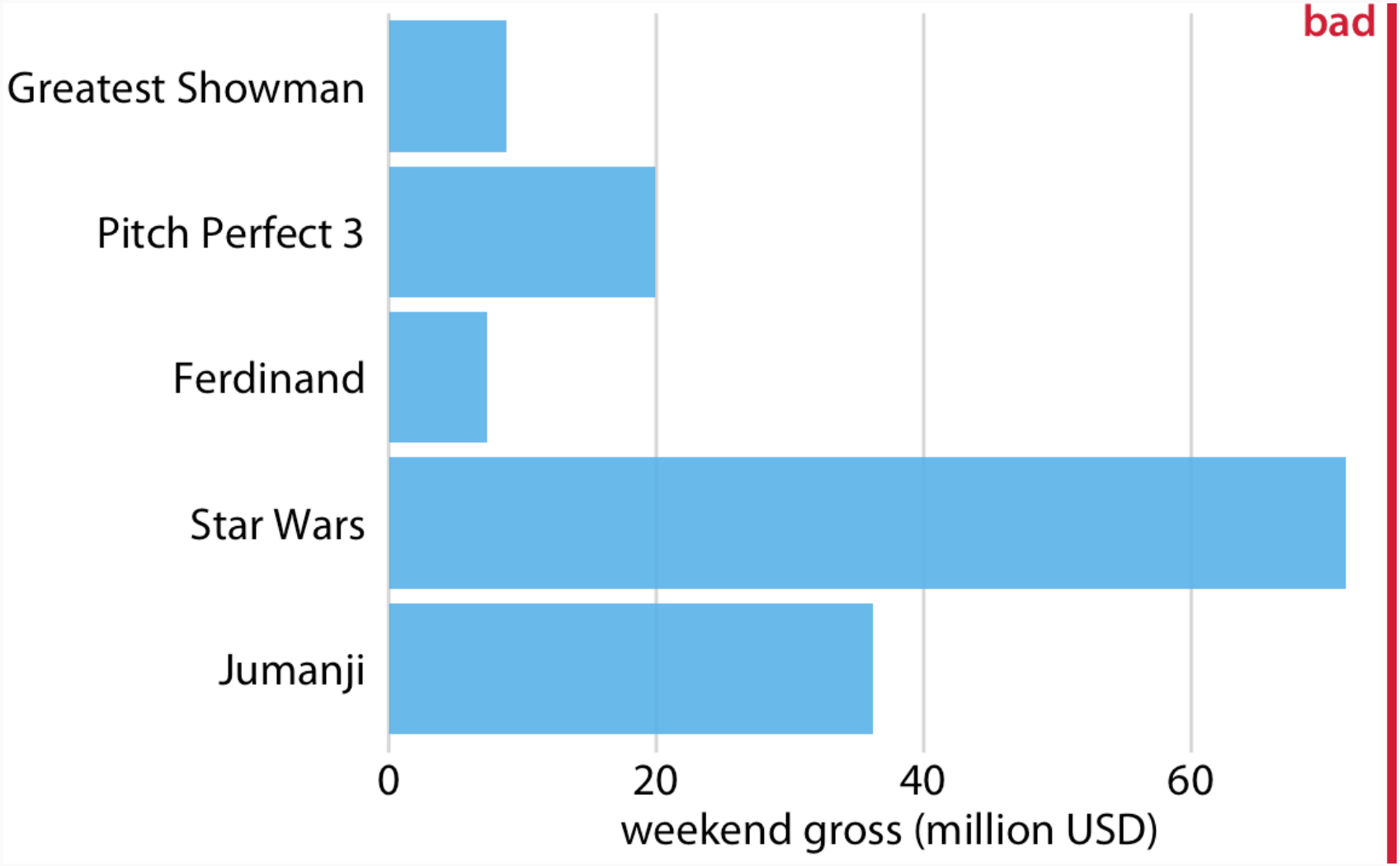


柱状图应该从 0 开始

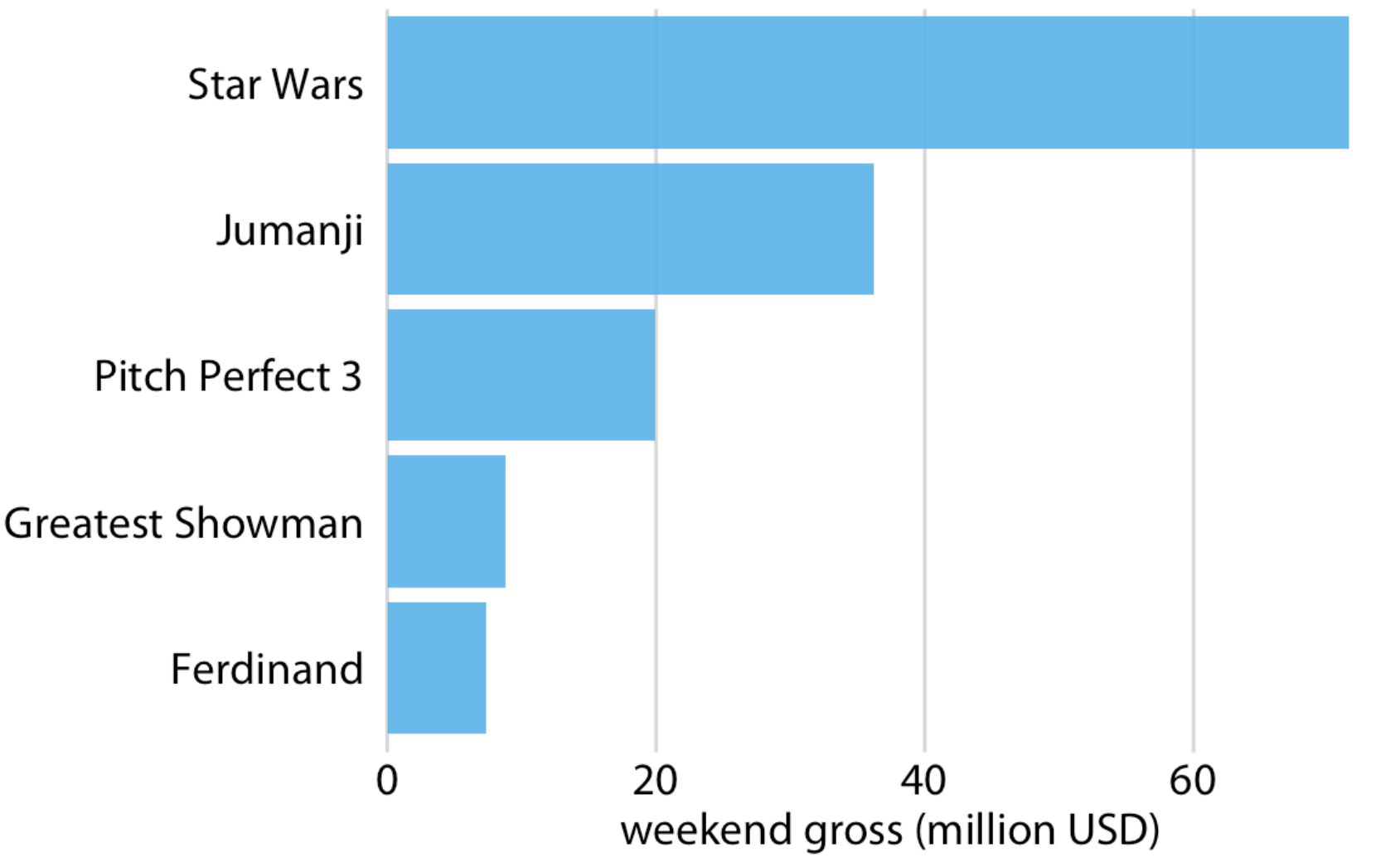
比例对等原则


$25,000
$20,000
$15,000
$10,000
$5,000
$0
-$5,000
5-year change in median income (USD)
Honolulu
Maui
Kalawao
Kauai
Hawaii
county

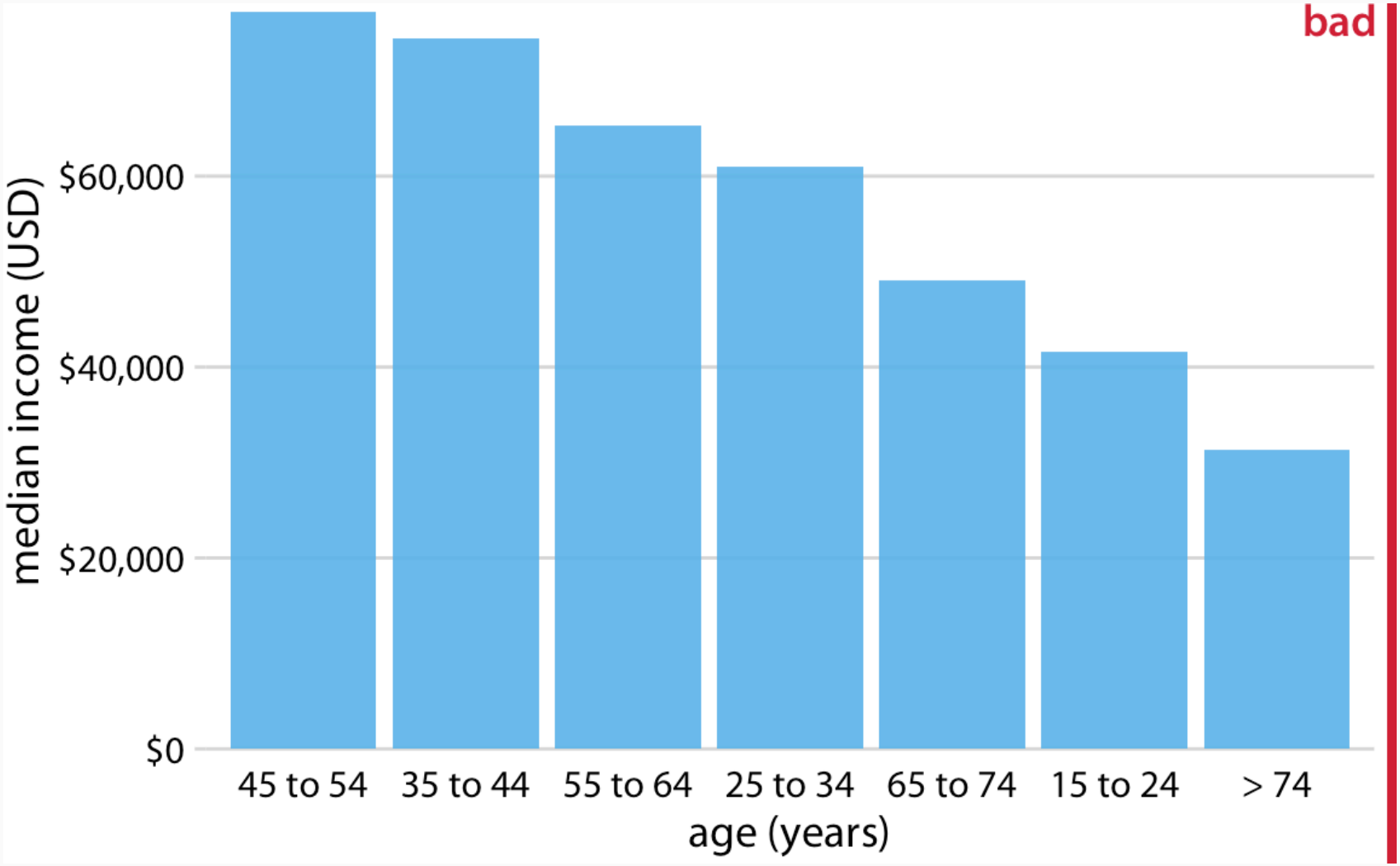
排序很关键


bad
Greatest Showman
Pitch Perfect 3
Ferdinand
Star Wars
Jumanji
0
20
40
60
weekend gross (million USD)

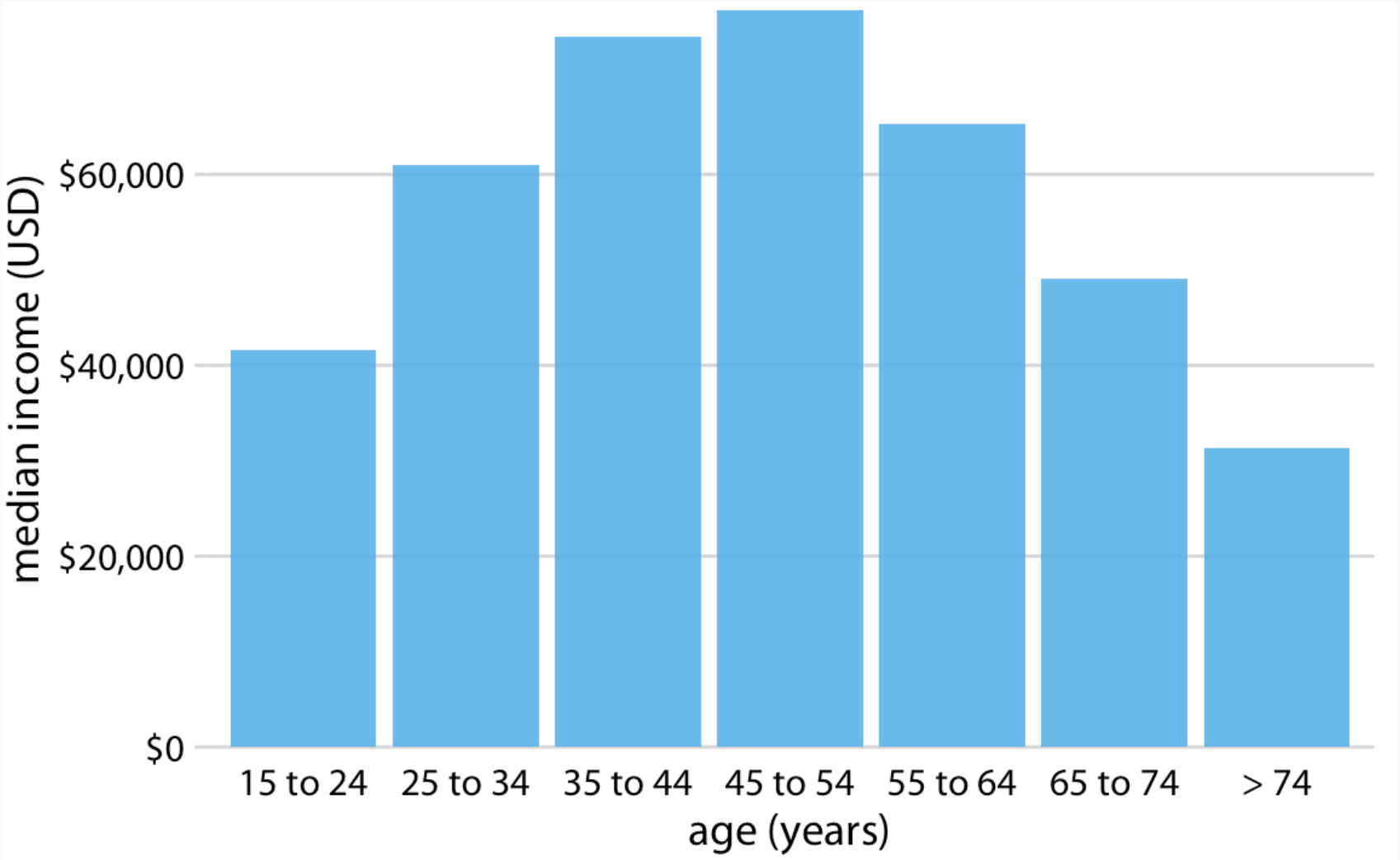
排序很关键


Star Wars
Jumanji
Pitch Perfect 3
Greatest Showman
Ferdinand
0
20
40
60
weekend gross (million USD)

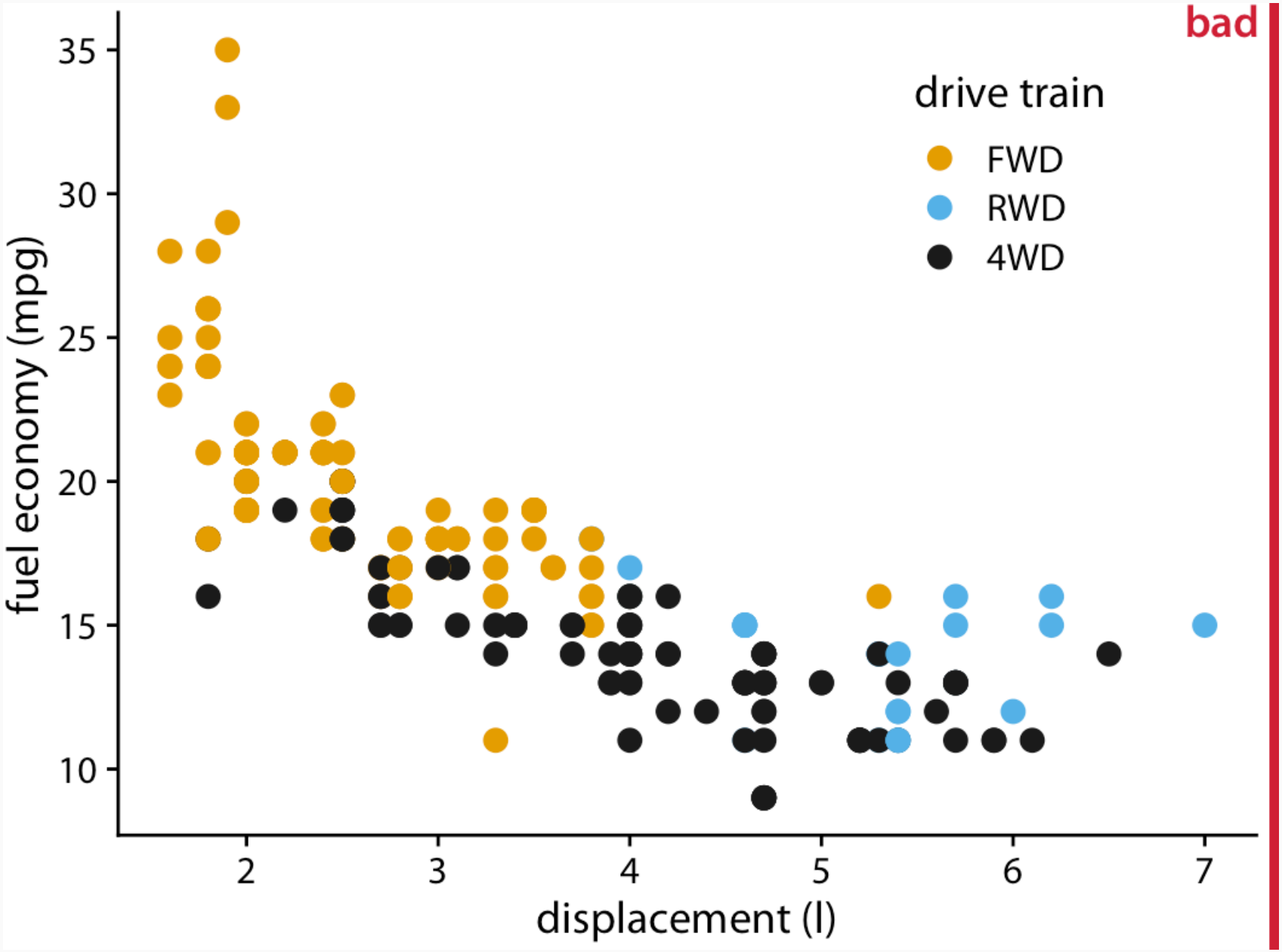
排序很关键


bad
median income (USD)
$60,000
$40,000
$20,000
$0
45 to 54
35 to 44
55 to 64
25 to 34
65 to 74
15 to 24
> 74
age (years)

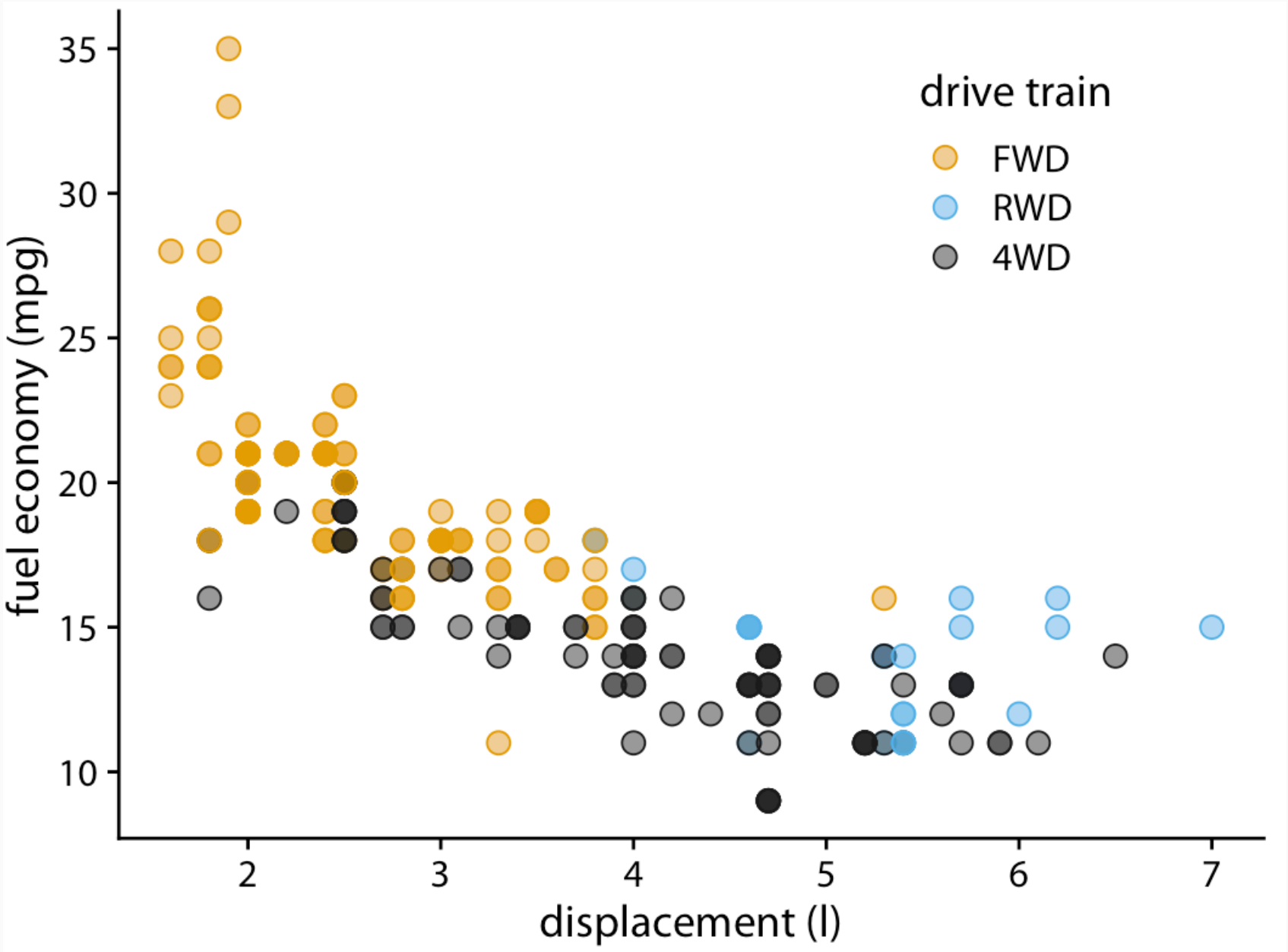
排序很关键


median income (USD)
$60,000
$40,000
$20,000
$0
15 to 24
25 to 34
35 to 44
45 to 54
55 to 64
65 to 74
> 74
age (years)

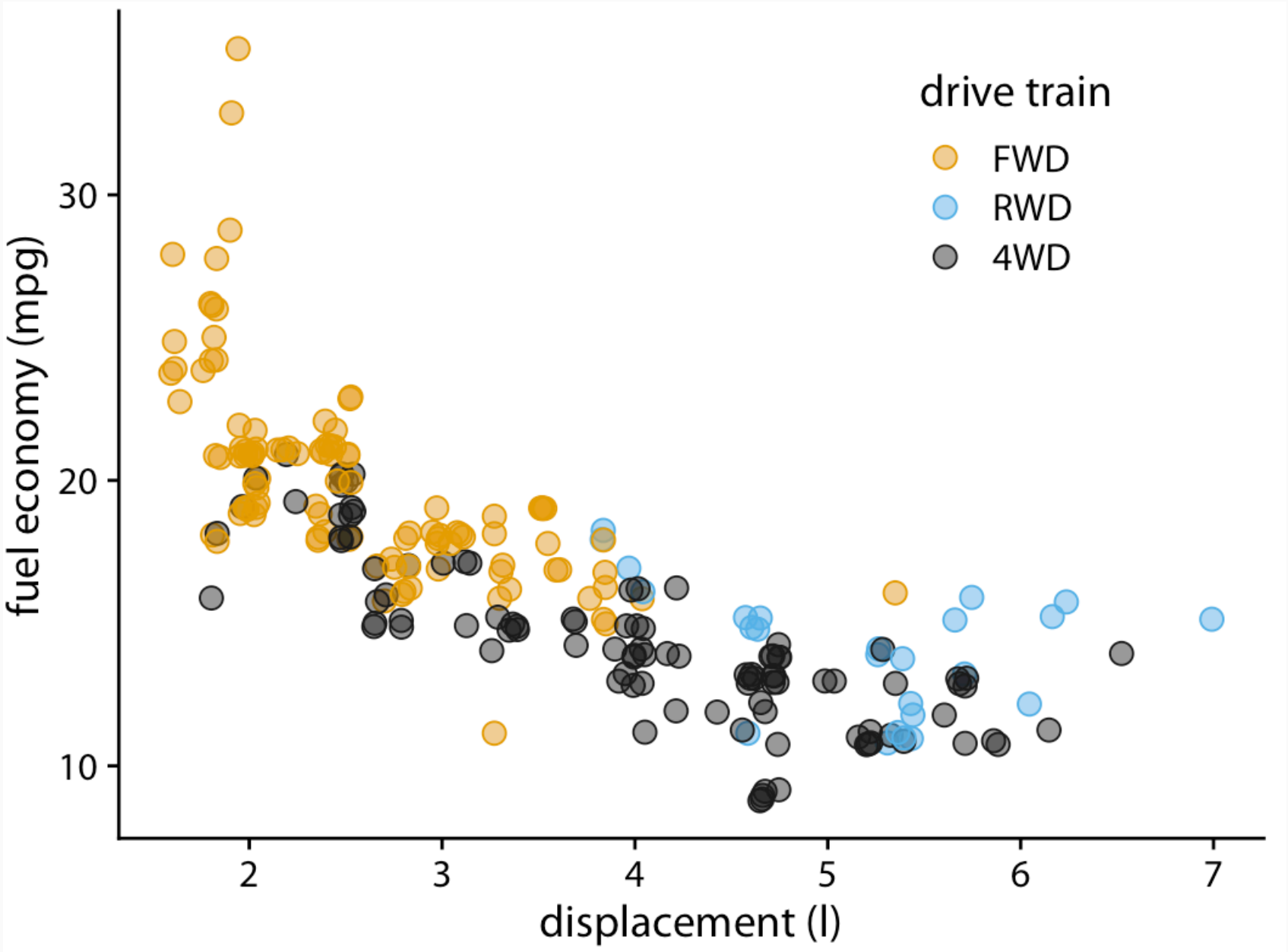
处理好重叠点


bad
drive train
FWD
RWD
4WD
fuel economy (mpg)
35
30
25
20
15
10
2
3
4
5
6
7
displacement (l)

处理好重叠点

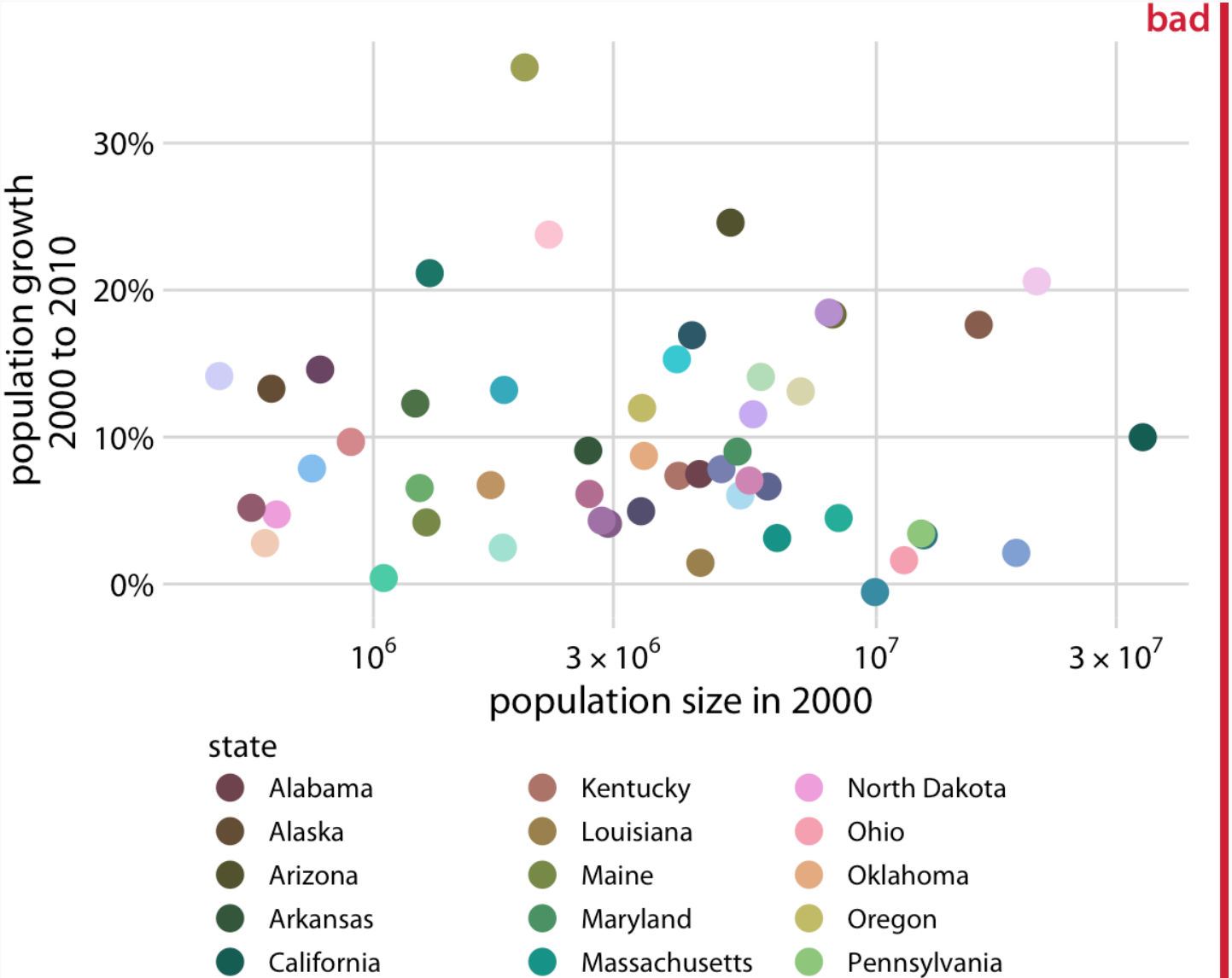

drive train
FWD
RWD
4WD
fuel economy (mpg)
35
30
25
20
15
10
2
3
4
5
6
7
displacement (l)

处理好重叠点


drive train
FWD
RWD
4WD
fuel economy (mpg)
30
20
10
2
3
4
5
6
7
displacement (l)

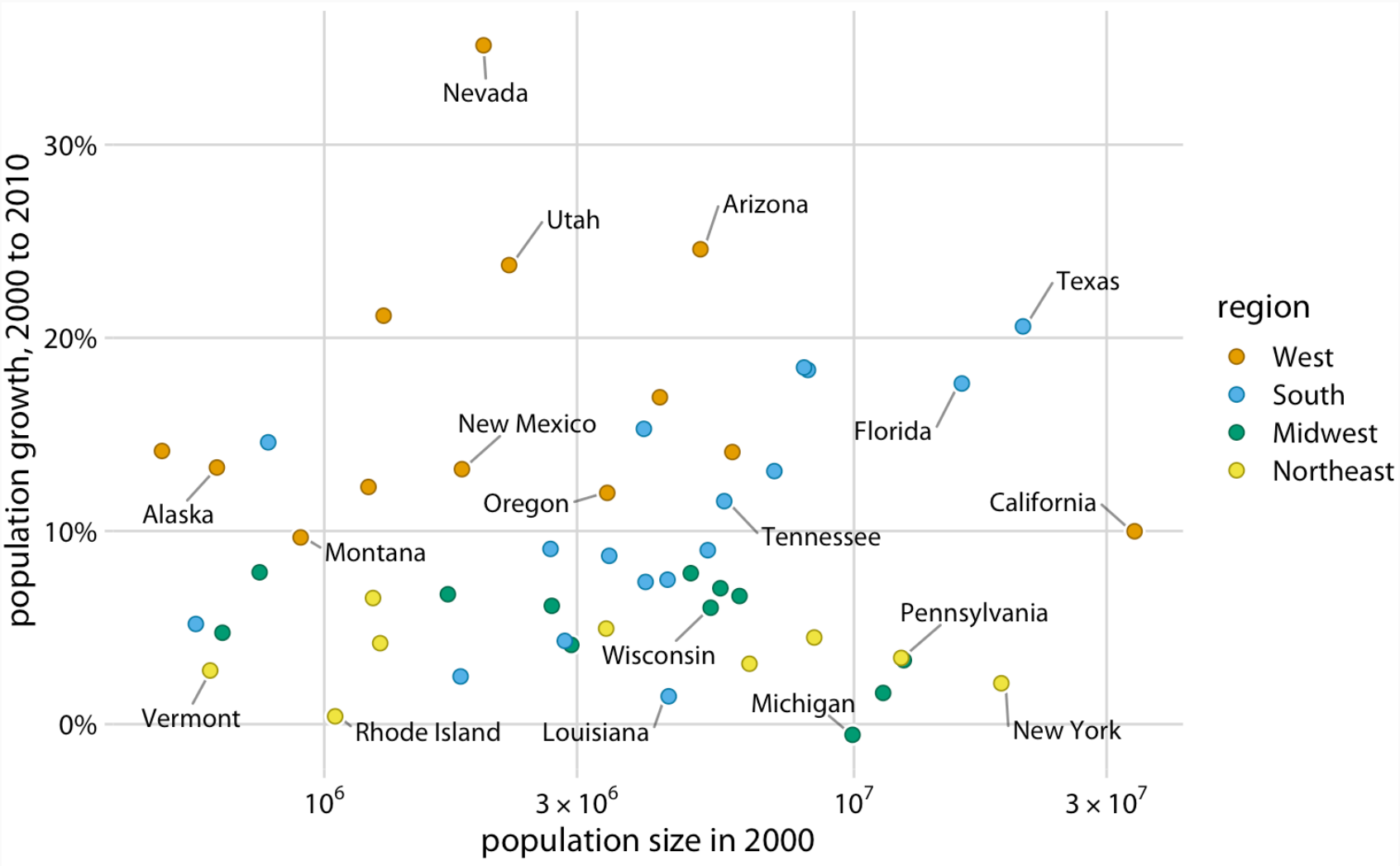
色彩的运用

颜色，可以提升可视化效果，也可毁掉本来很好的图

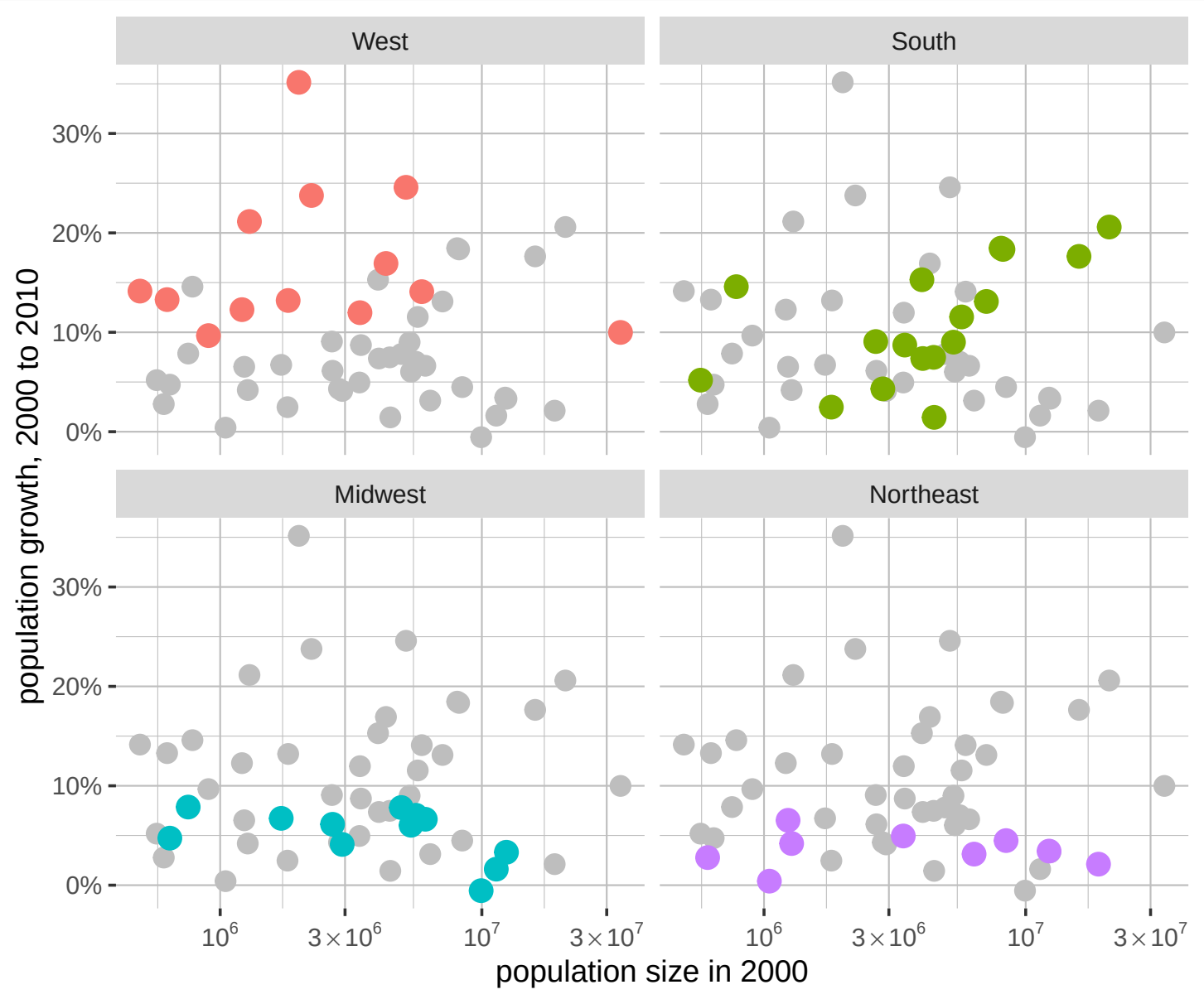


色彩的运用

颜色分组，同时增加标注

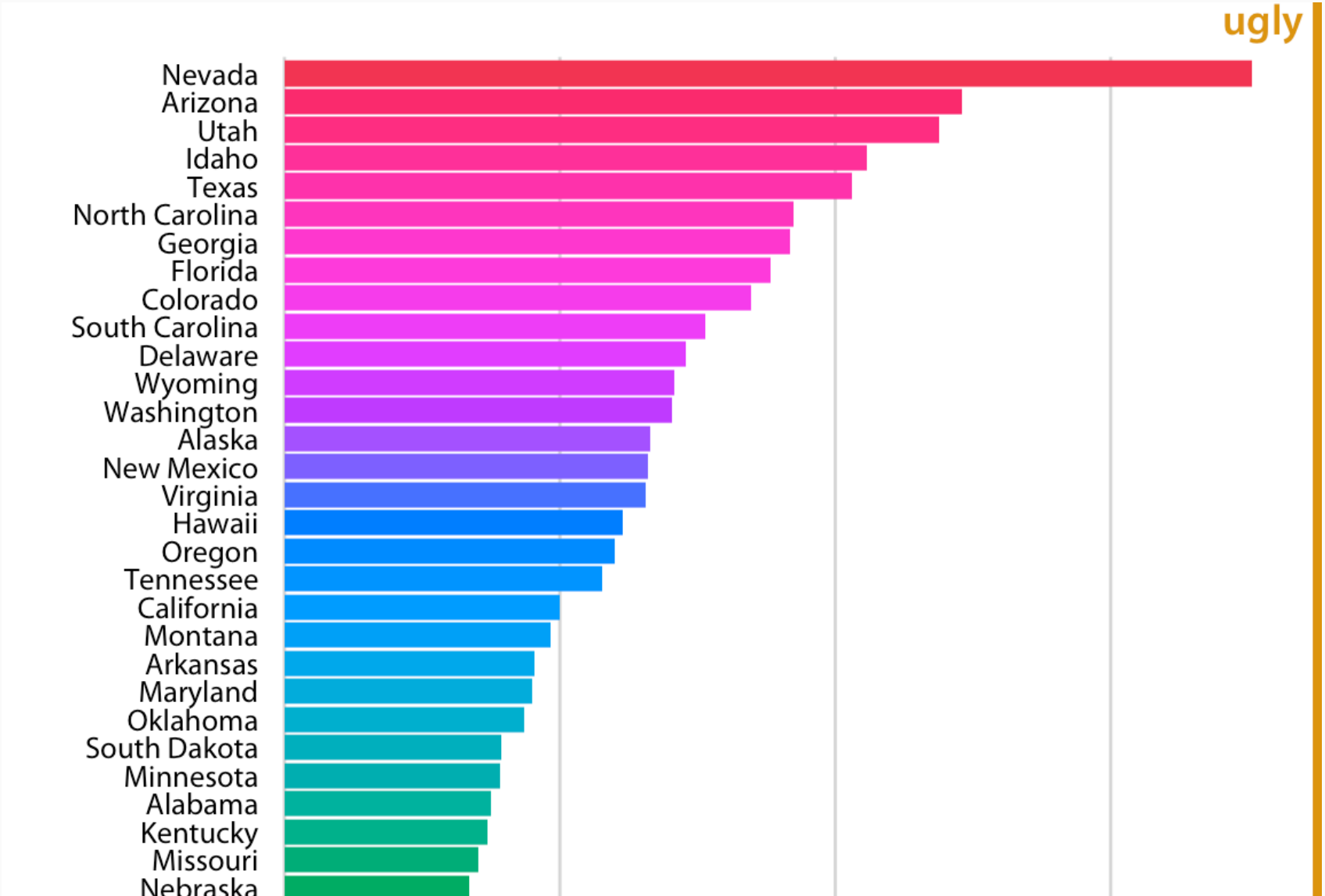


色彩的运用

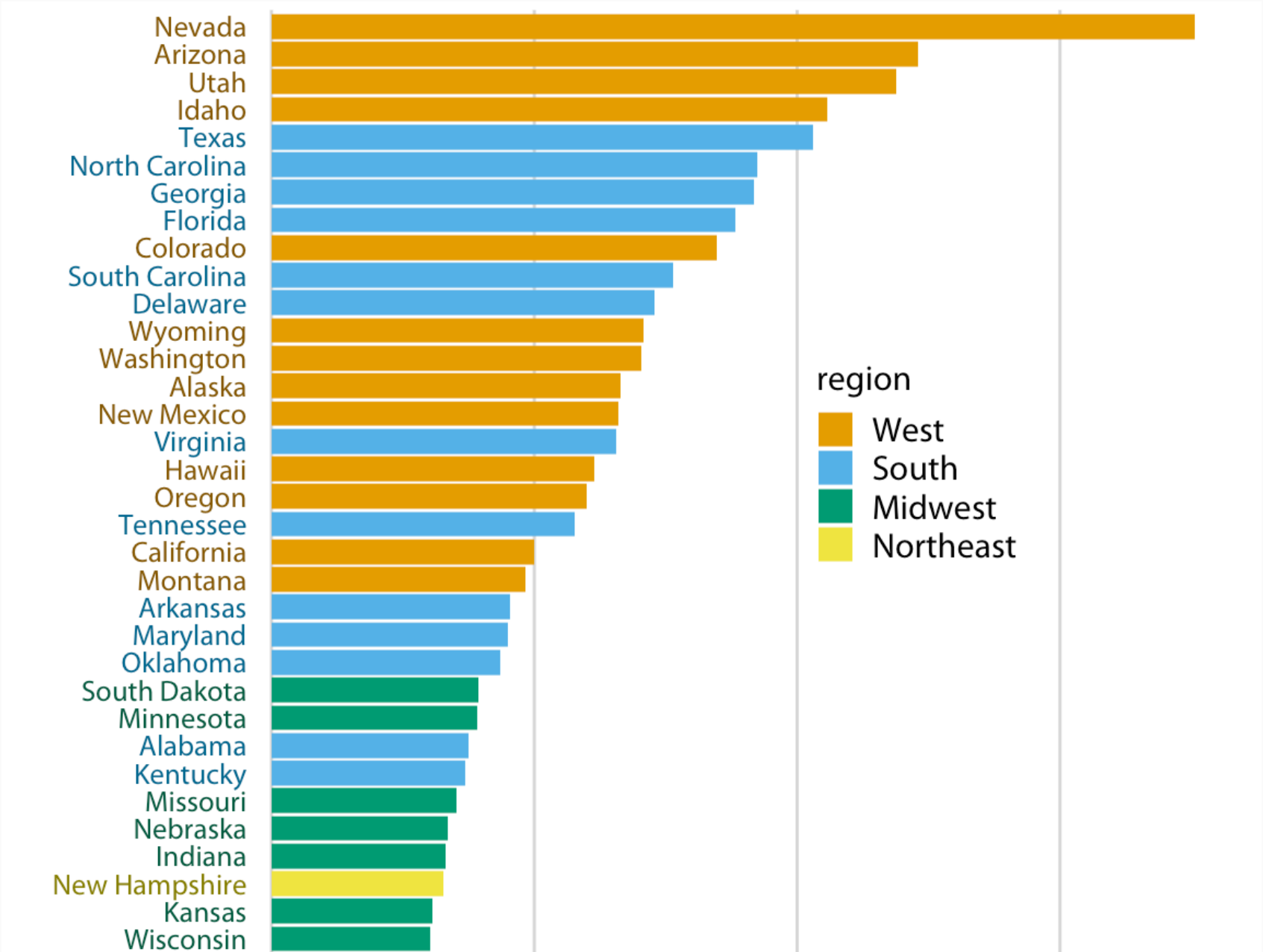

West
South
Midwest
Northeast
30%
20%
10%
0%
population growth, 2000 to 2010
10^6
3×10^6
10^7
3×10^7
population size in 2000

色彩的运用

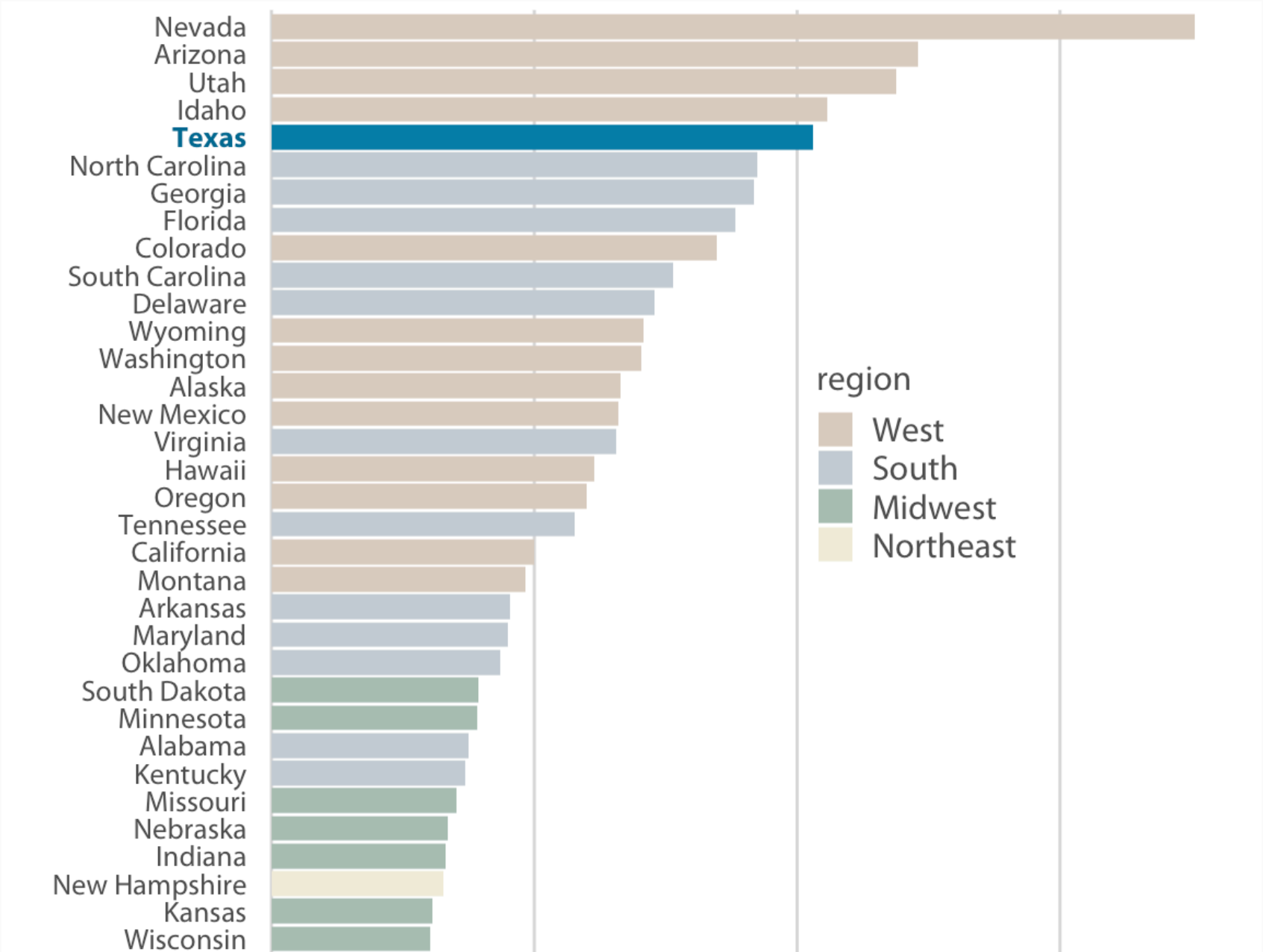
为了颜色而颜色，但一点信息量也没有



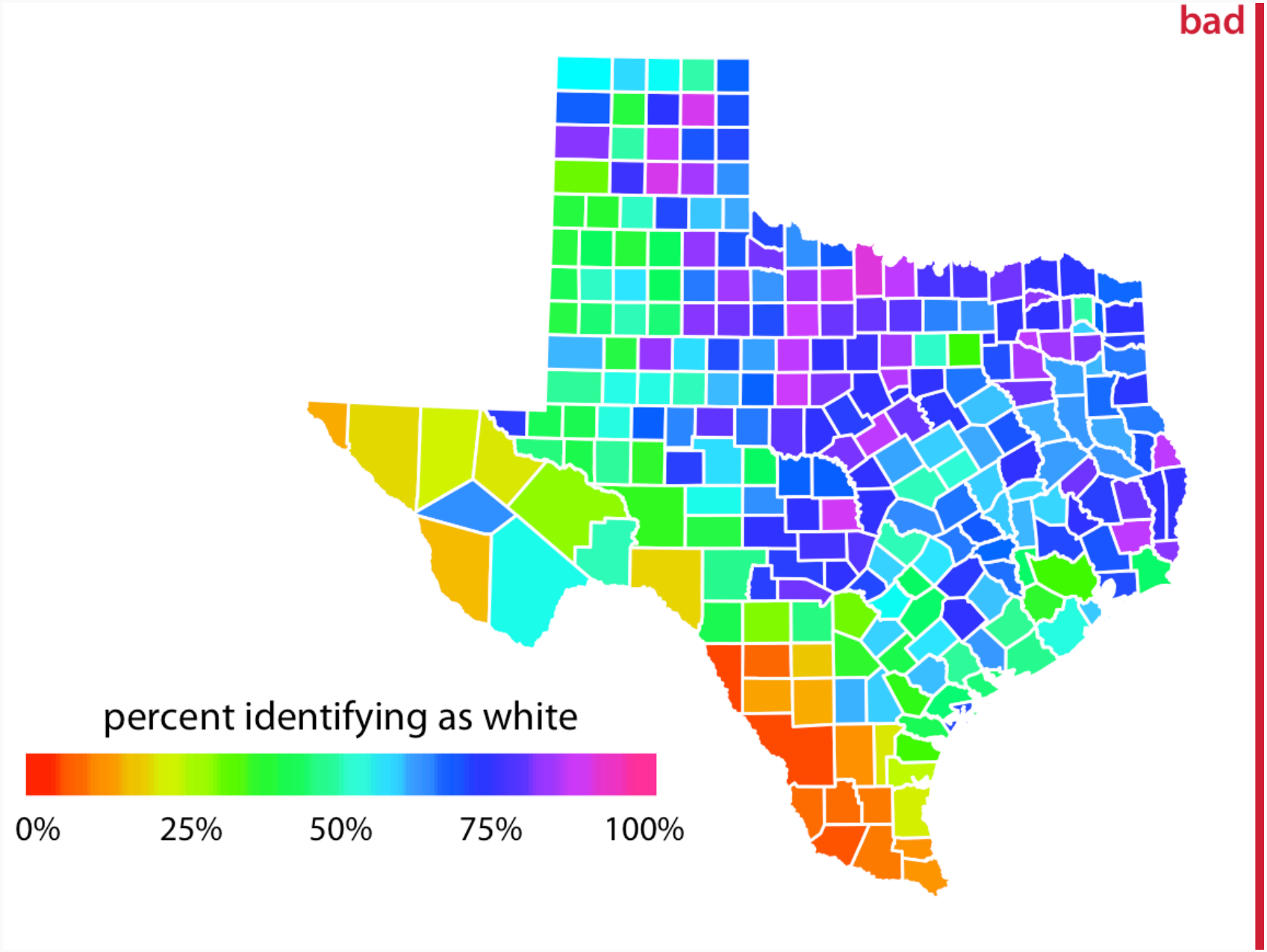
色彩的运用


Nevada
Arizona
Utah
Idaho
Texas
North Carolina
Georgia
Florida
Colorado
South Carolina
Delaware
Wyoming
Washington
Alaska
New Mexico
Virginia
Hawaii
Oregon
Tennessee
California
Montana
Arkansas
Maryland
Oklahoma
South Dakota
Minnesota
Alabama
Kentucky
Missouri
Nebraska
Indiana
New Hampshire
Kansas
Wisconsin
region
West
South
Midwest
Northeast

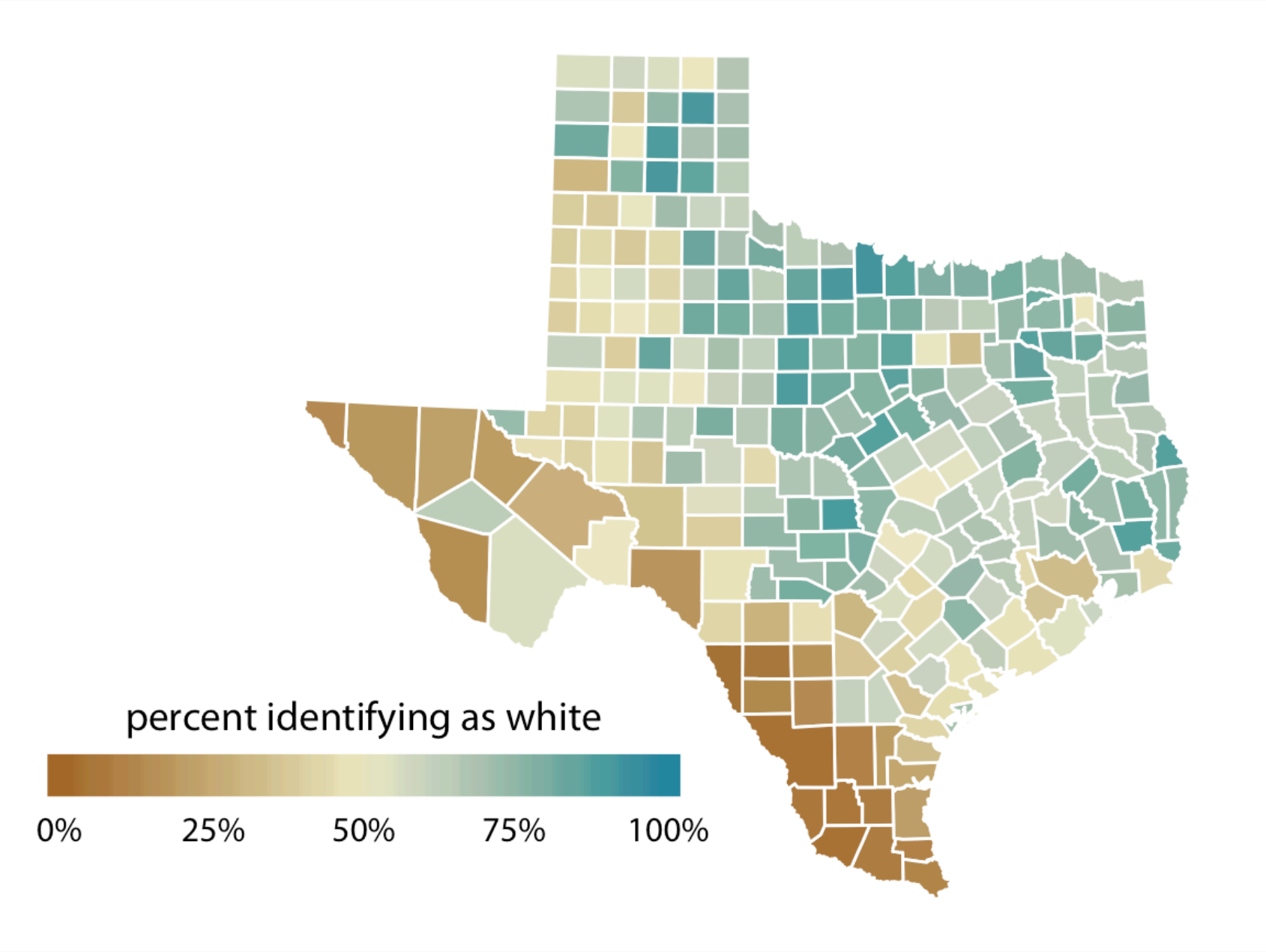
色彩的运用


Nevada
Arizona
Utah
Idaho
Texas
North Carolina
Georgia
Florida
Colorado
South Carolina
Delaware
Wyoming
Washington
Alaska
New Mexico
Virginia
Hawaii
Oregon
Tennessee
California
Montana
Arkansas
Maryland
Oklahoma
South Dakota
Minnesota
Alabama
Kentucky
Missouri
Nebraska
Indiana
New Hampshire
Kansas
Wisconsin
region
West
South
Midwest
Northeast

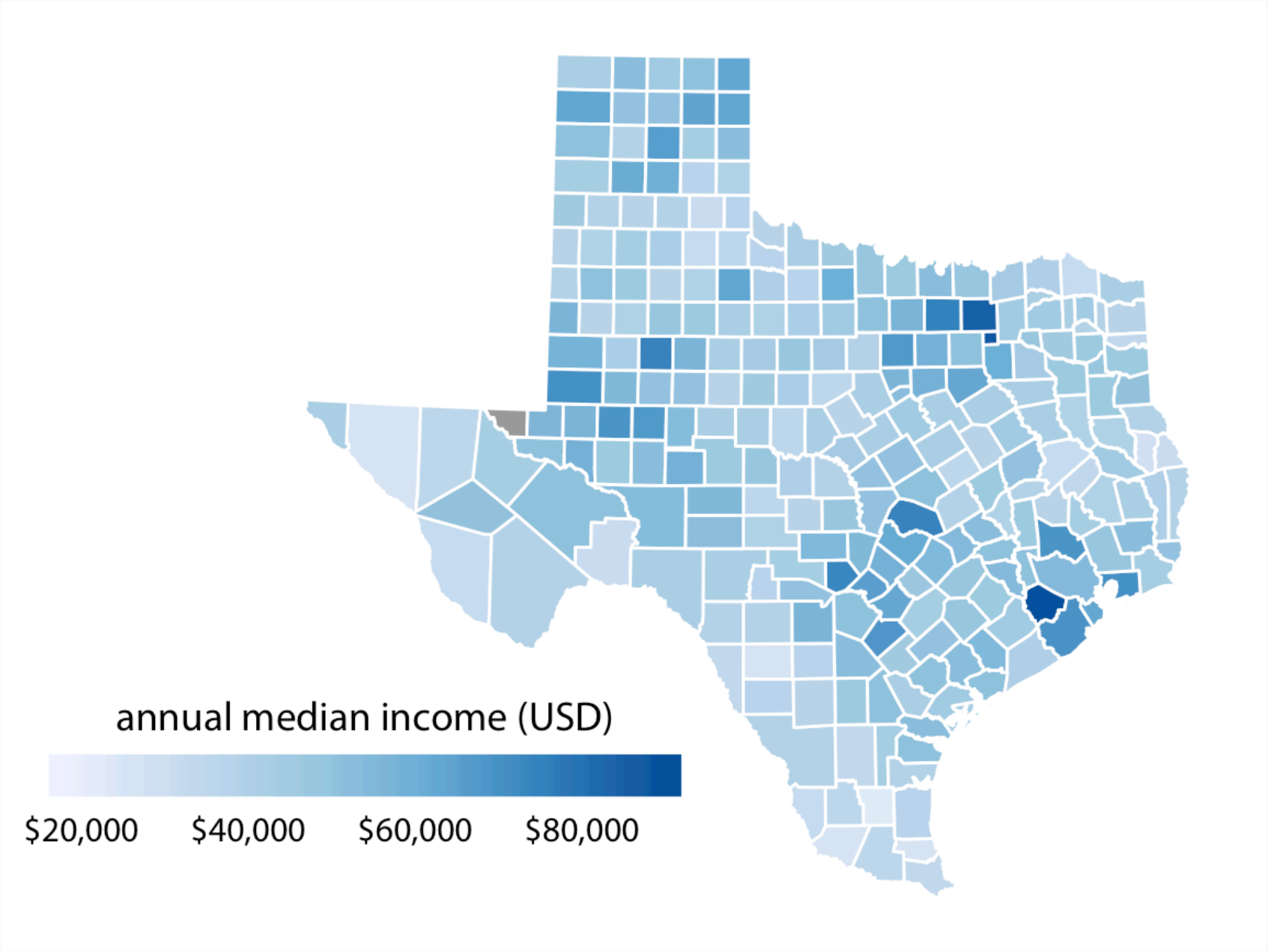
色彩的运用



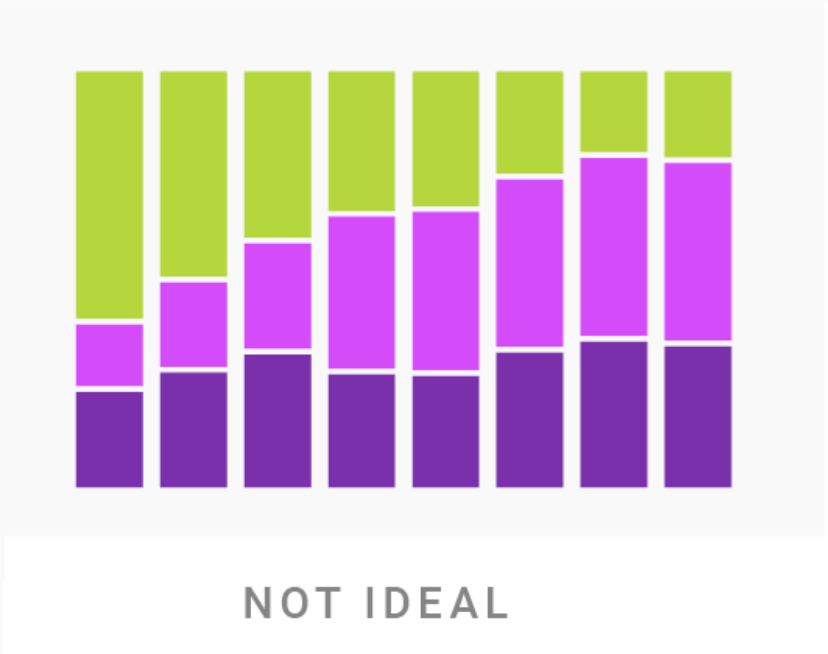
色彩的运用

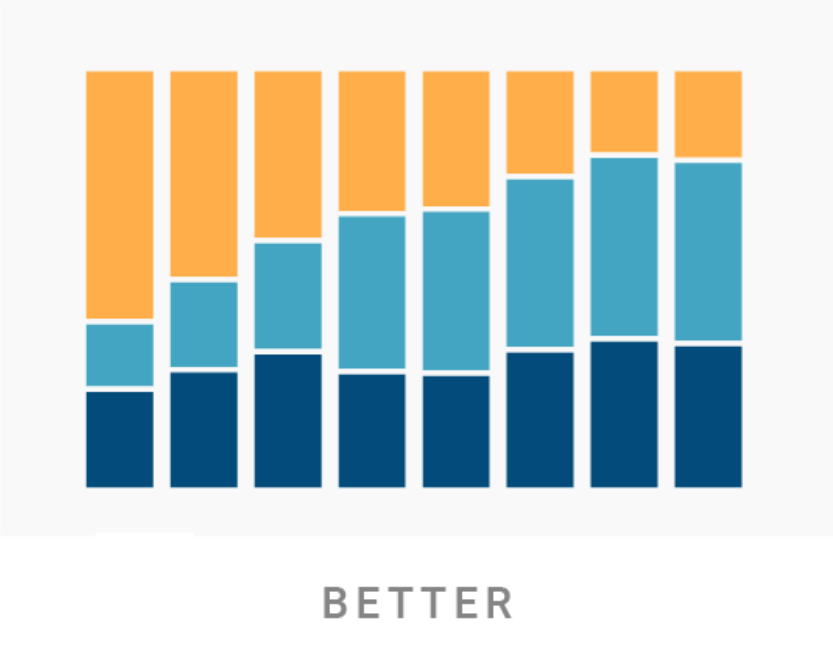


色彩的运用

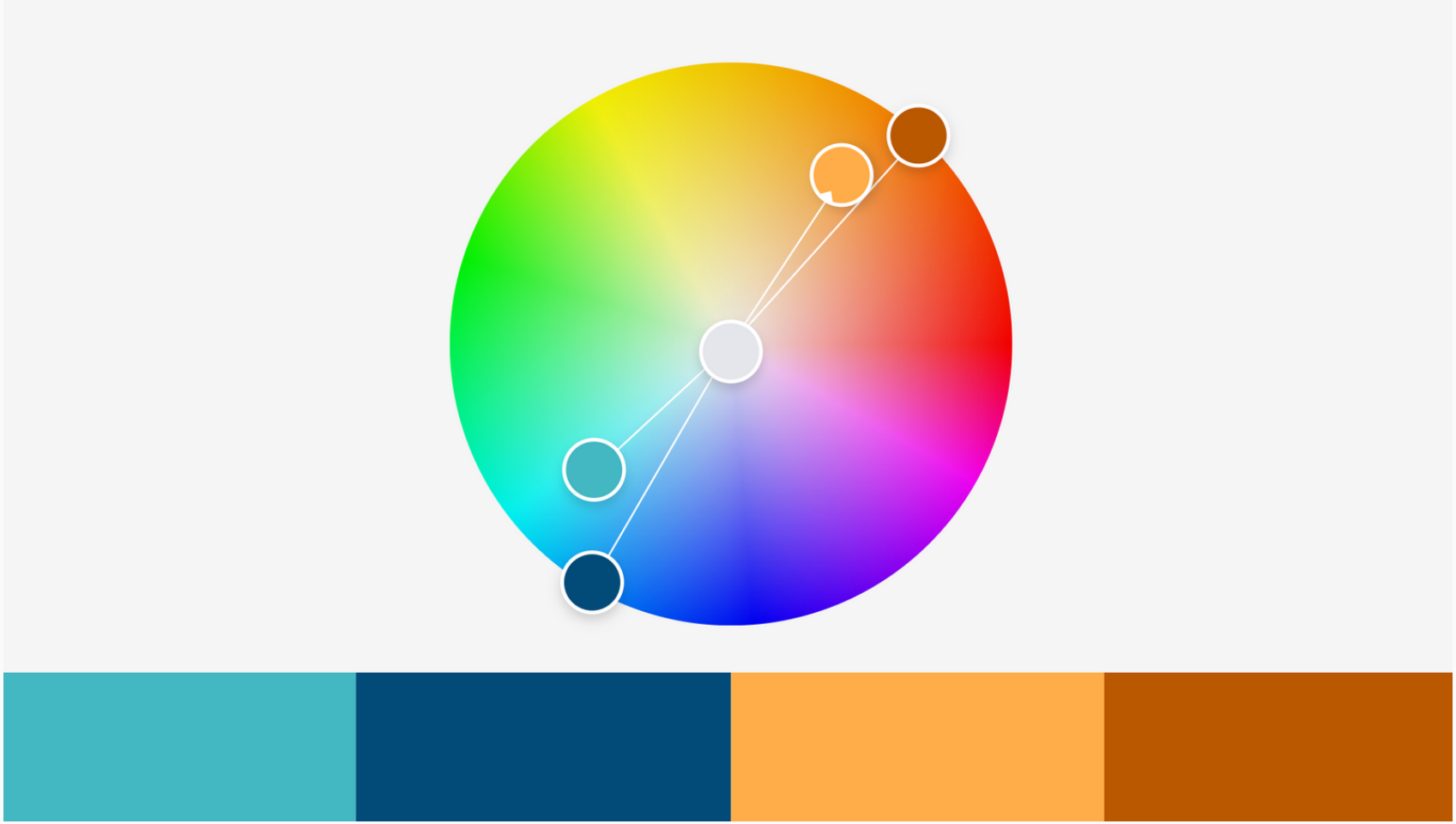


色彩的运用





色彩的运用

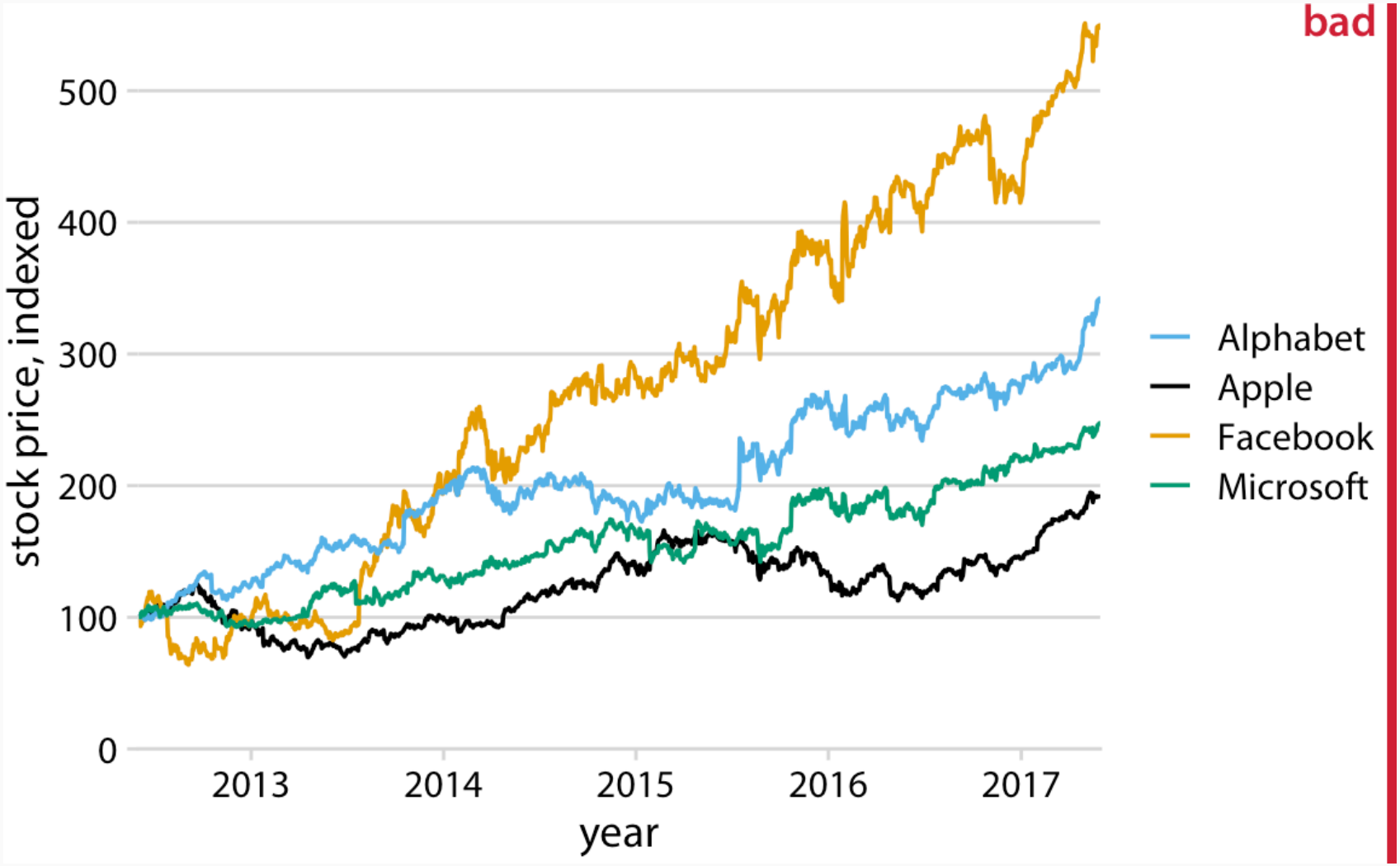


Adobe color

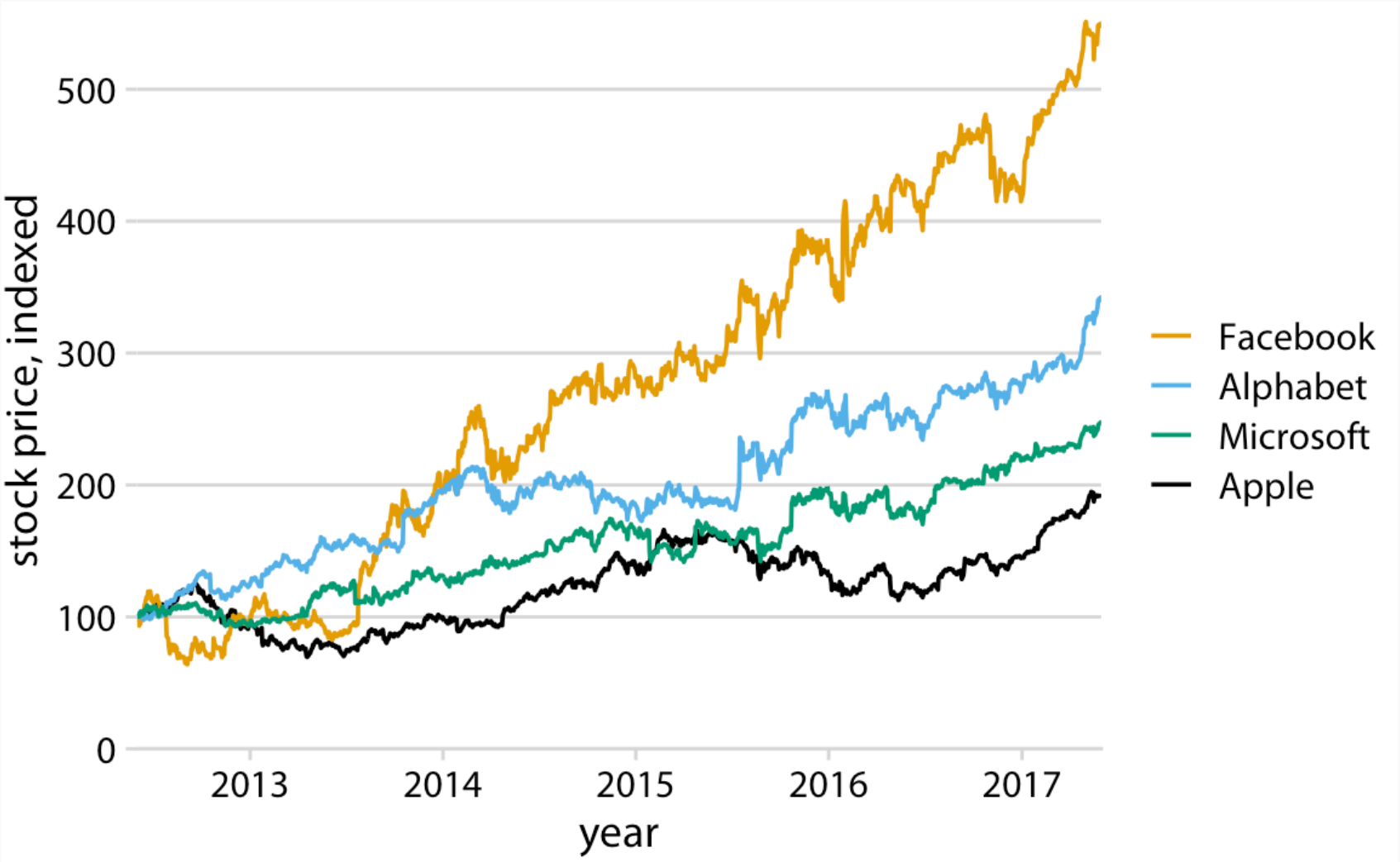
色彩的运用

怎么样把配色放到我的图中呢?

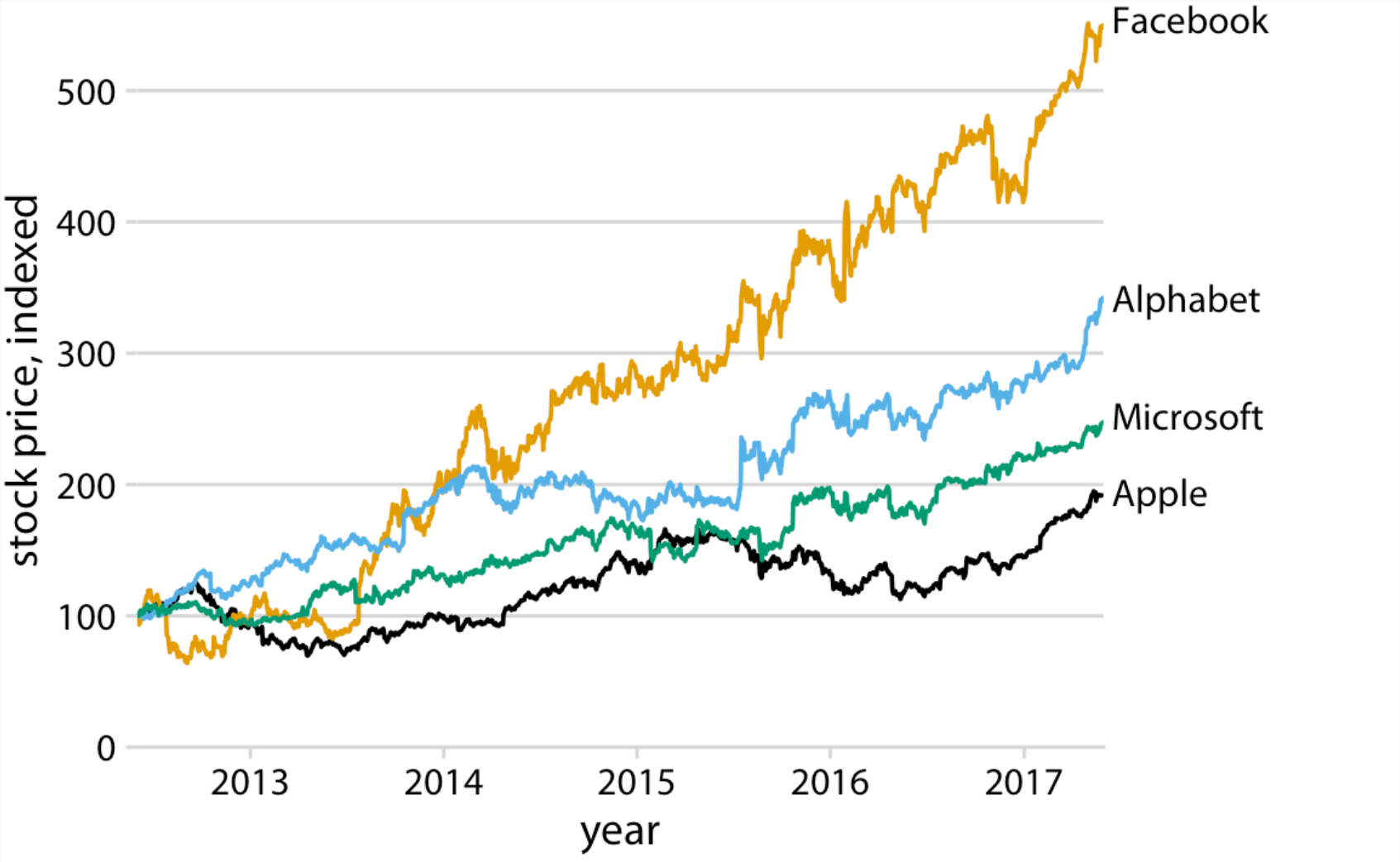
必要的标注


bad
500
400
300
200
100
0
stock price, indexed
2013
2014
2015
2016
2017
year
Alphabet
Apple
Facebook
Microsoft

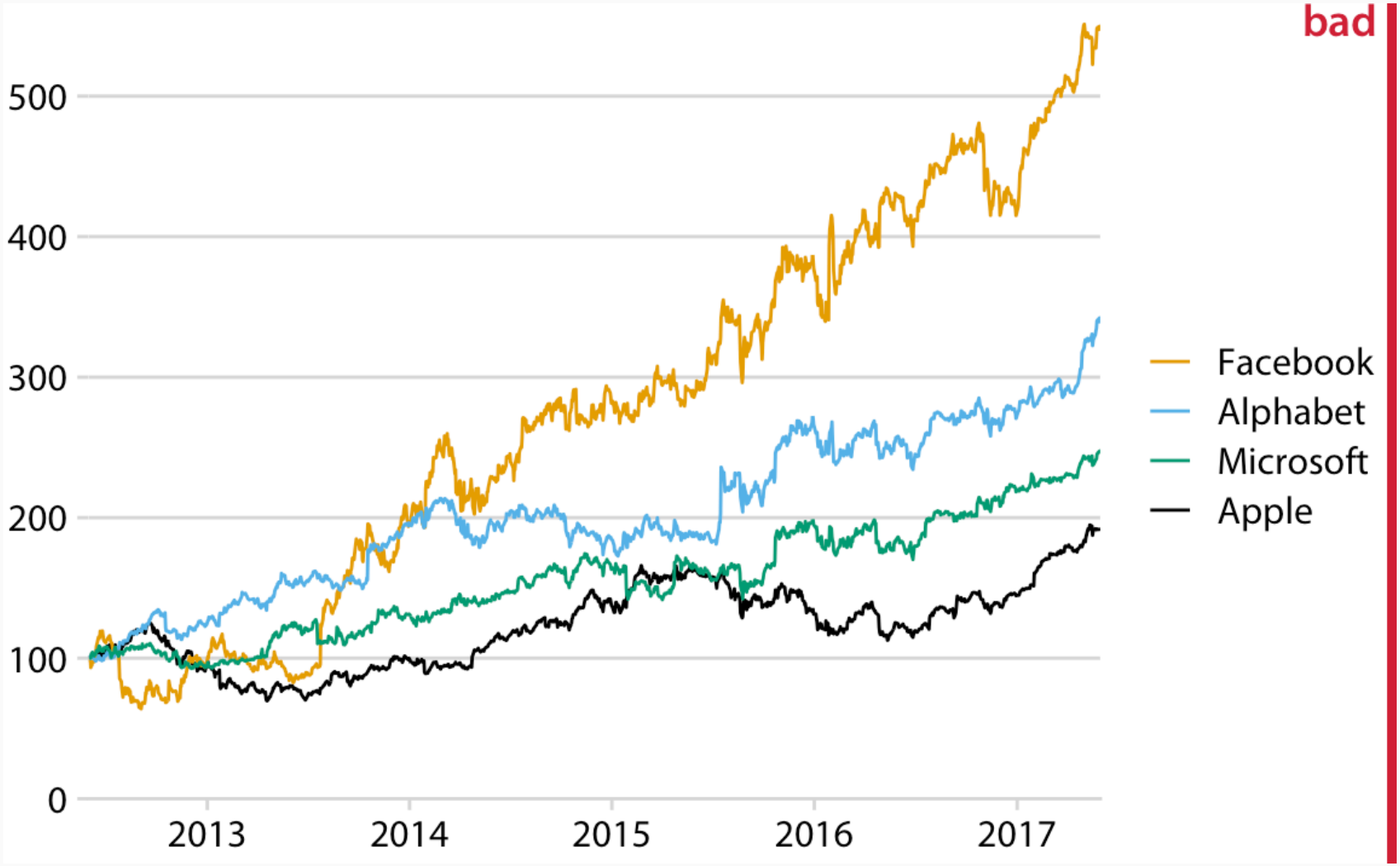
必要的标注


500
400
300
200
100
0
stock price, indexed
2013
2014
2015
2016
2017
year
Facebook
Alphabet
Microsoft
Apple

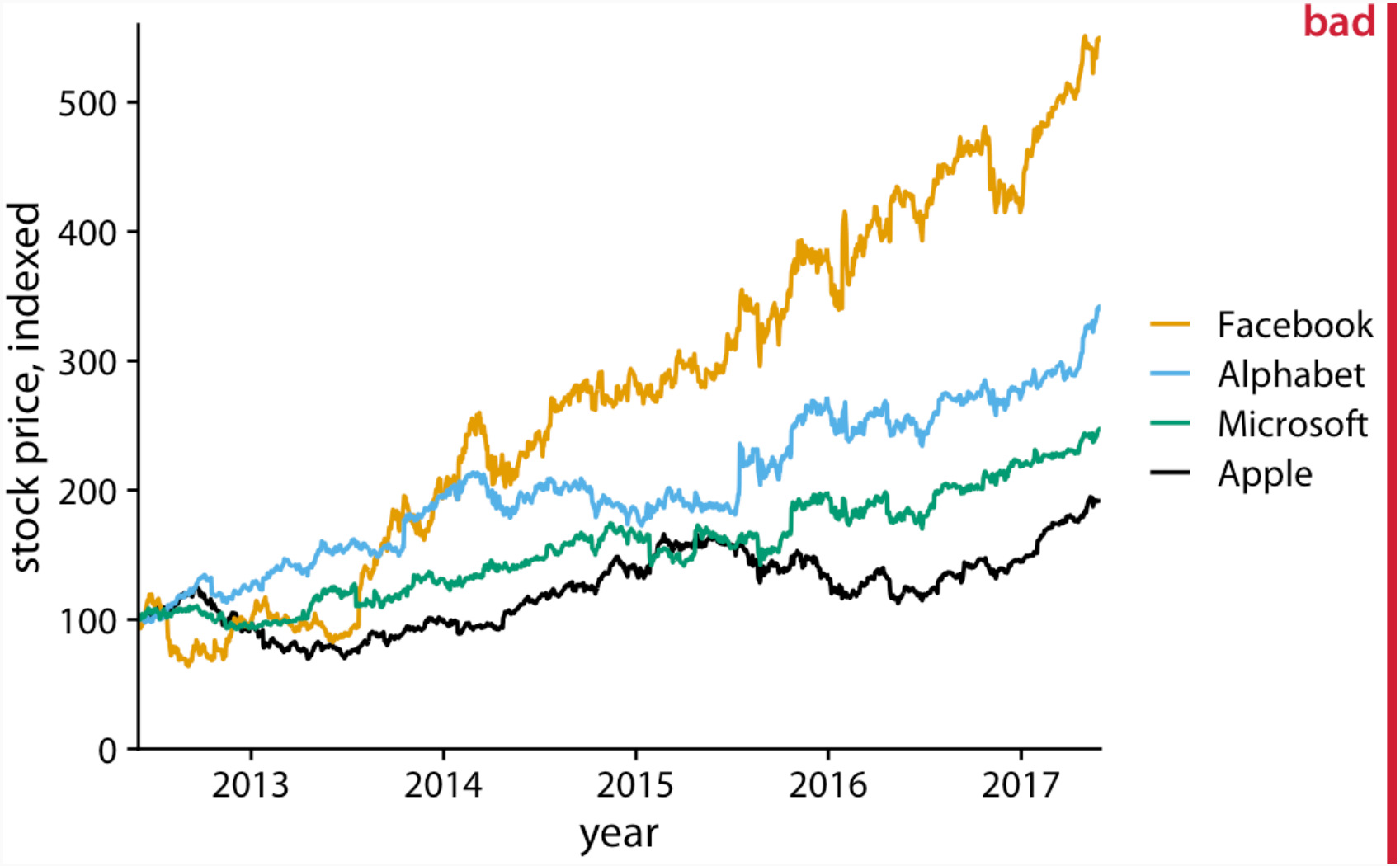
必要的标注


Facebook
Alphabet
Microsoft
Apple
500
400
300
200
100
0
stock price, indexed
2013
2014
2015
2016
2017
year

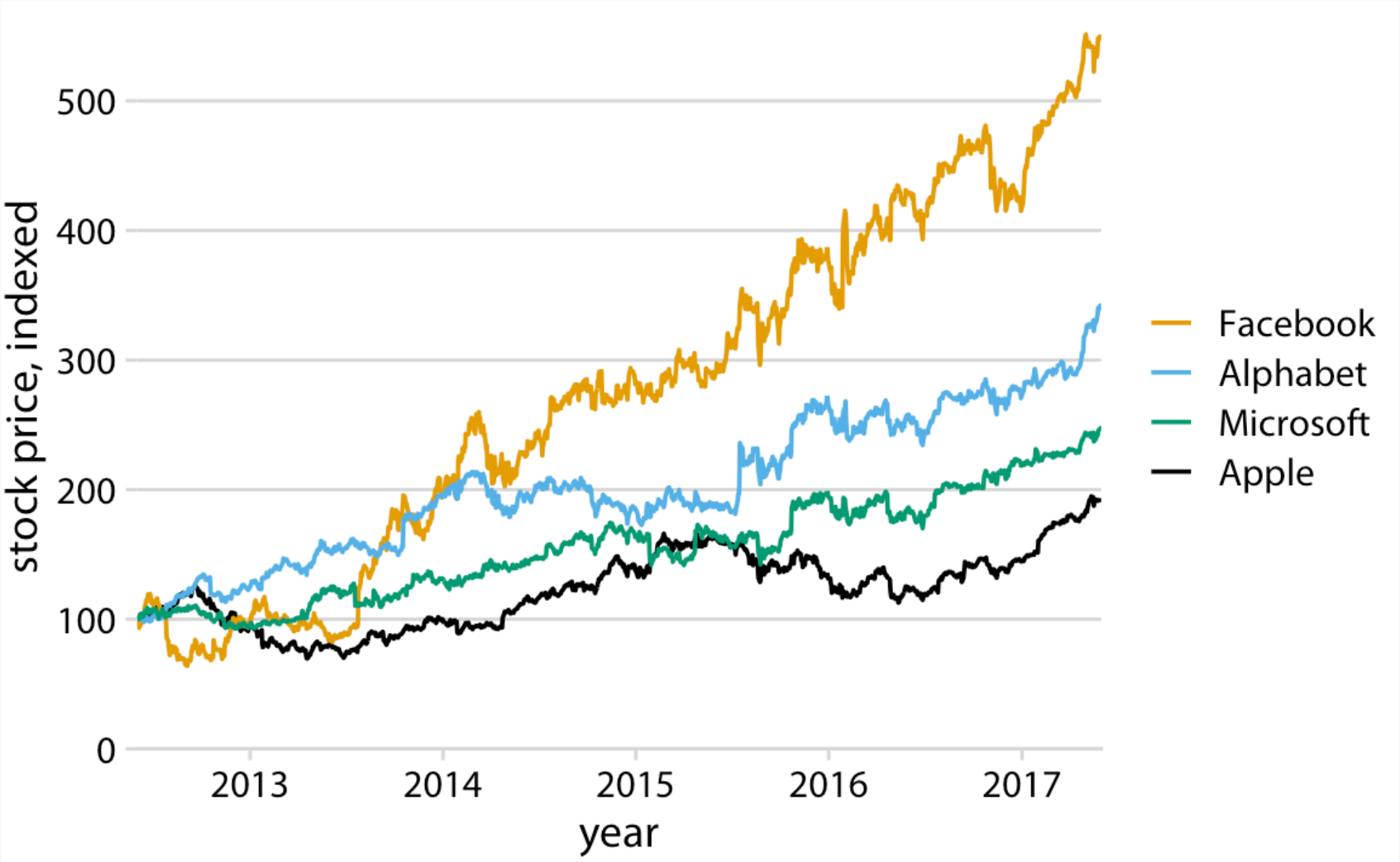
必要的标注


bad
500
400
300
200
100
0
2013
2014
2015
2016
2017
Facebook
Alphabet
Microsoft
Apple

必要的标注


bad
stock price, indexed
500
400
300
200
100
0
2013
2014
2015
2016
2017
year
Facebook
Alphabet
Microsoft
Apple

必要的标注

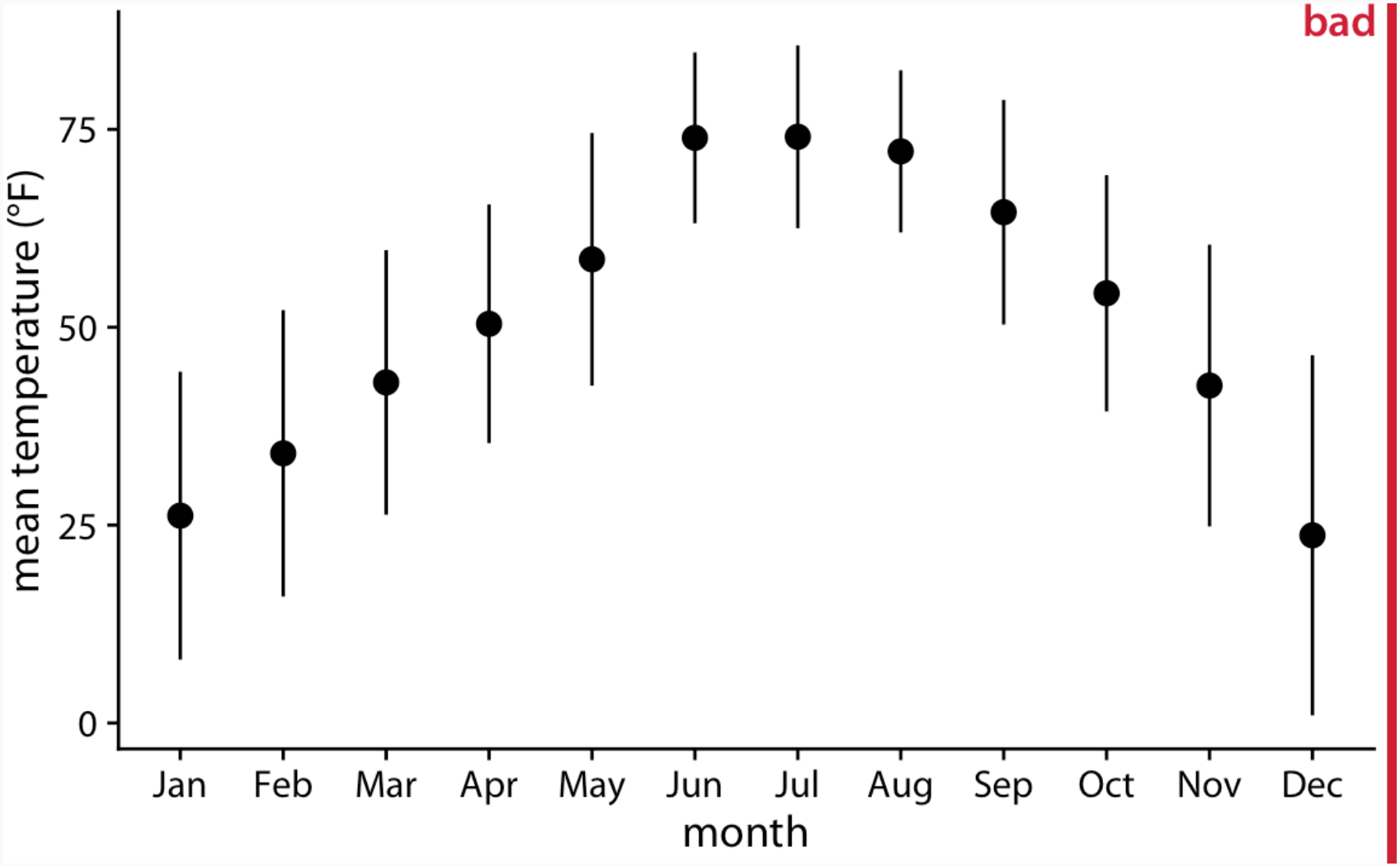

500
400
300
200
100
0
stock price, indexed
2013
2014
2015
2016
2017
year
Facebook
Alphabet
Microsoft
Apple

分布图的优缺点

表 1: 美国林肯市，2016 年气温数据

Month	month_short	Mean Temperature [F]
January	Jan	24
January	Jan	23
January	Jan	23
January	Jan	17
January	Jan	29
January	Jan	33
January	Jan	30

points-errorbars



箱线图

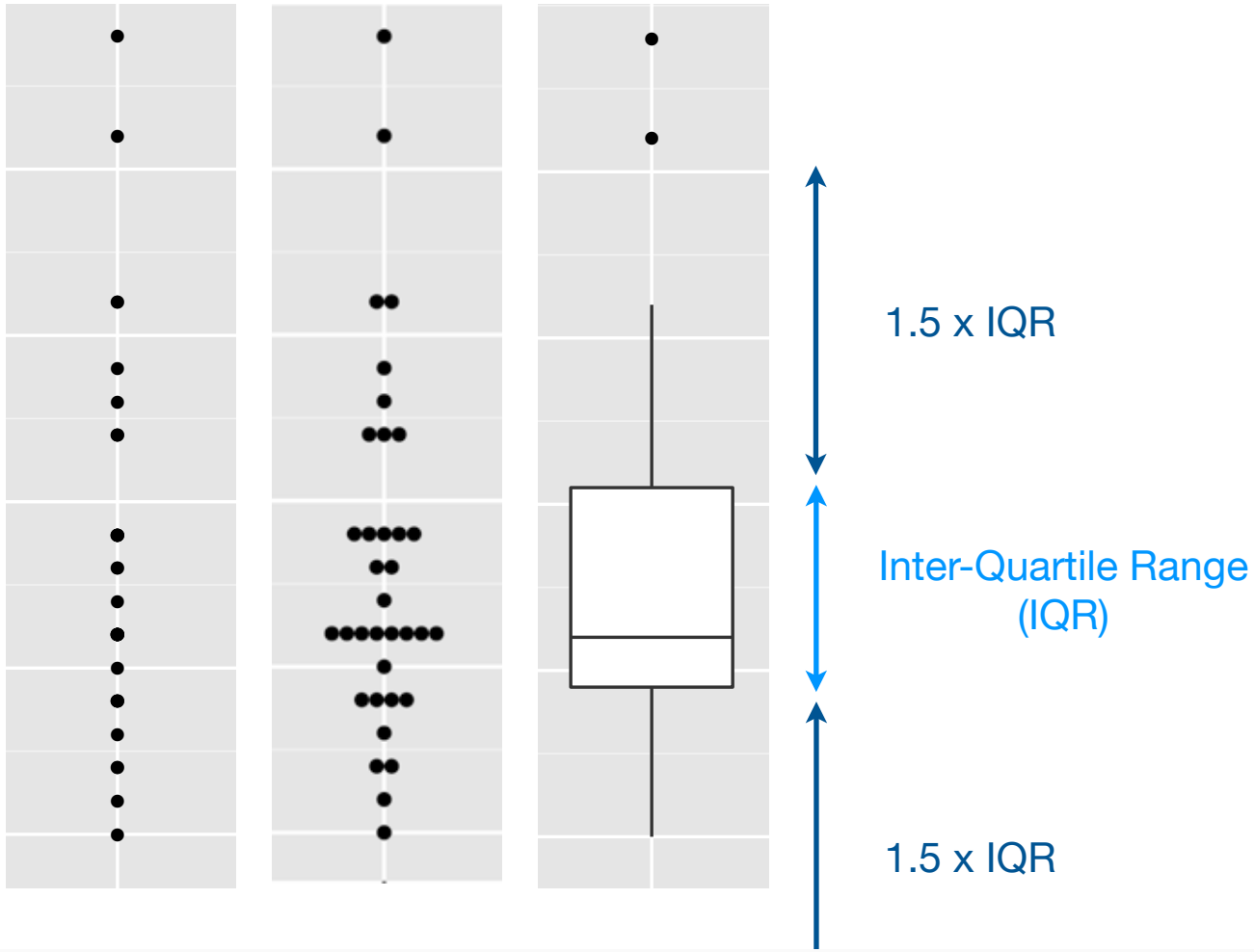
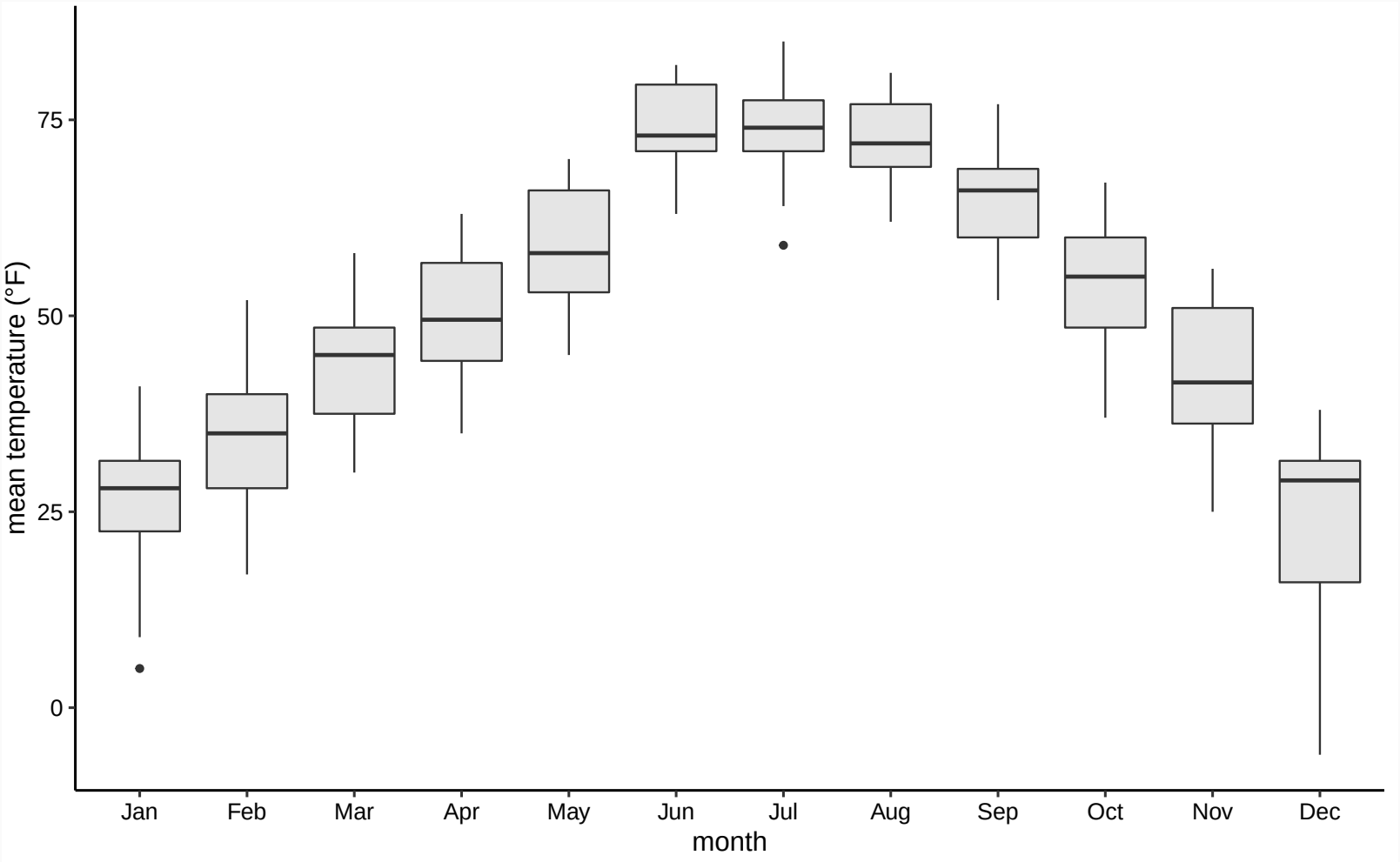
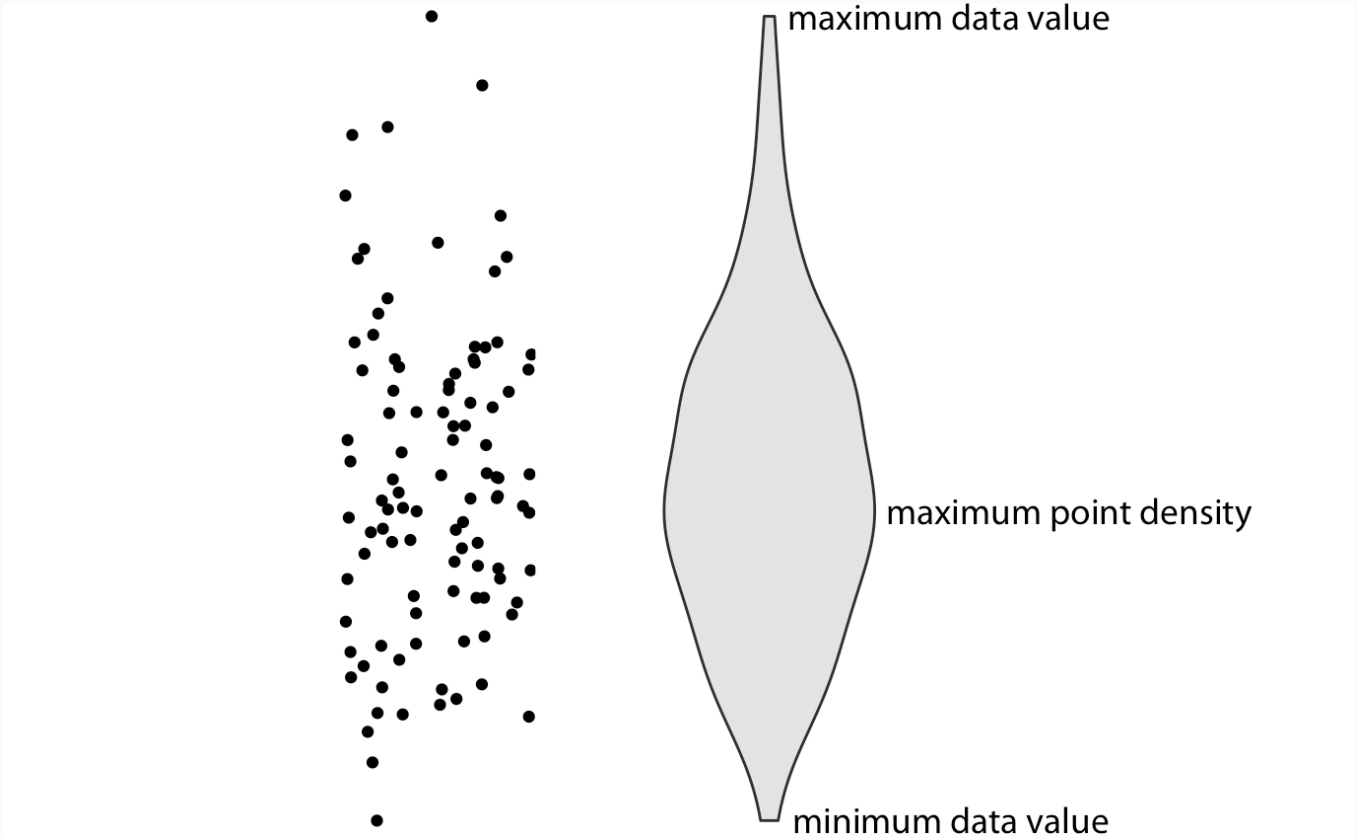


图 1: 箱线图示意图

箱线图


mean temperature (°F)
75
50
25
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month

小提琴图


maximum data value
maximum point density
minimum data value

小提琴图的优势

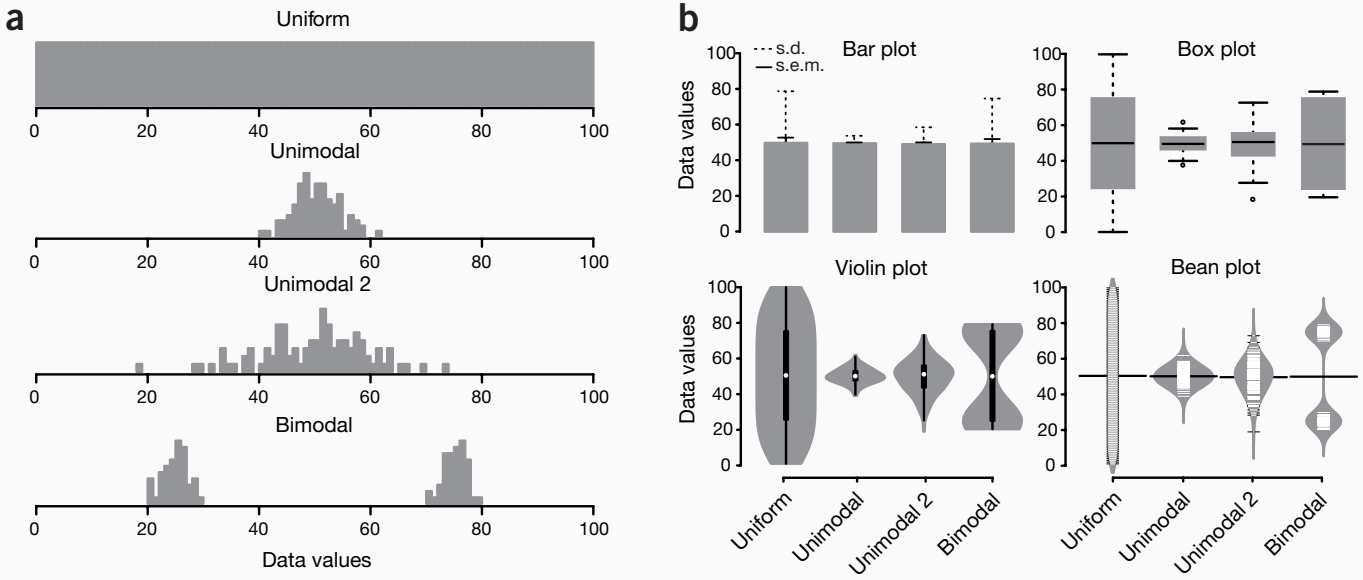
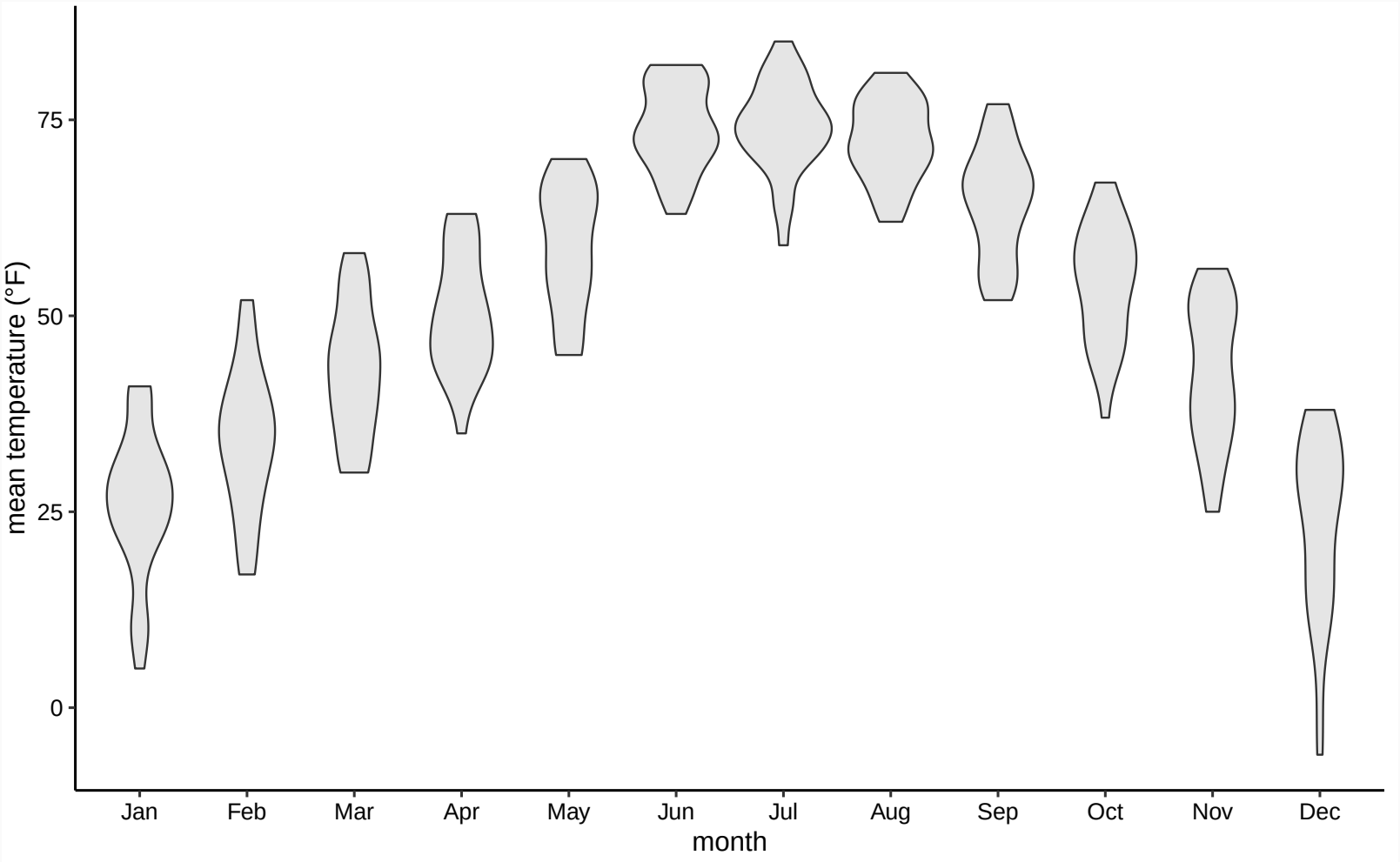


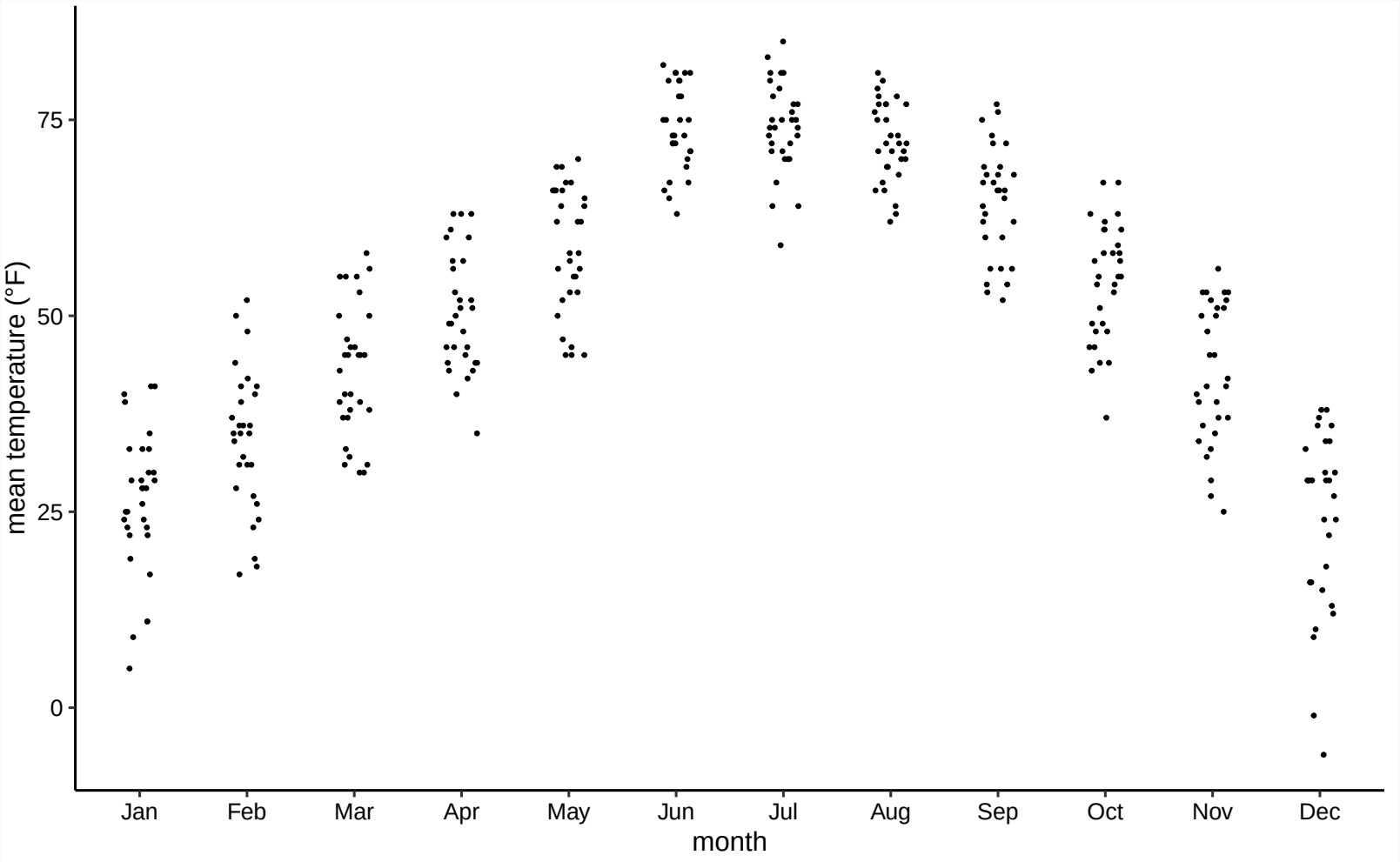
Figure 1 | Data visualization with box plots. (**a**) Hypothetical sample data sets of 100 data points each that are uniform, unimodal with one of two different variances or bimodal. Simple bar plot representations and statistical parameters may obscure such different data distributions. (**b**) Comparison of data visualization methods. Bar plots typically represent only the mean and s.d. or s.e.m. Box plots visualize the five-number summary of a data set (minimum, lower quartile, median, upper quartile and maximum). Violin and bean plots represent the actual distribution of the individual data sets.

图 2: nature methods, vol.11, no.2, 2014

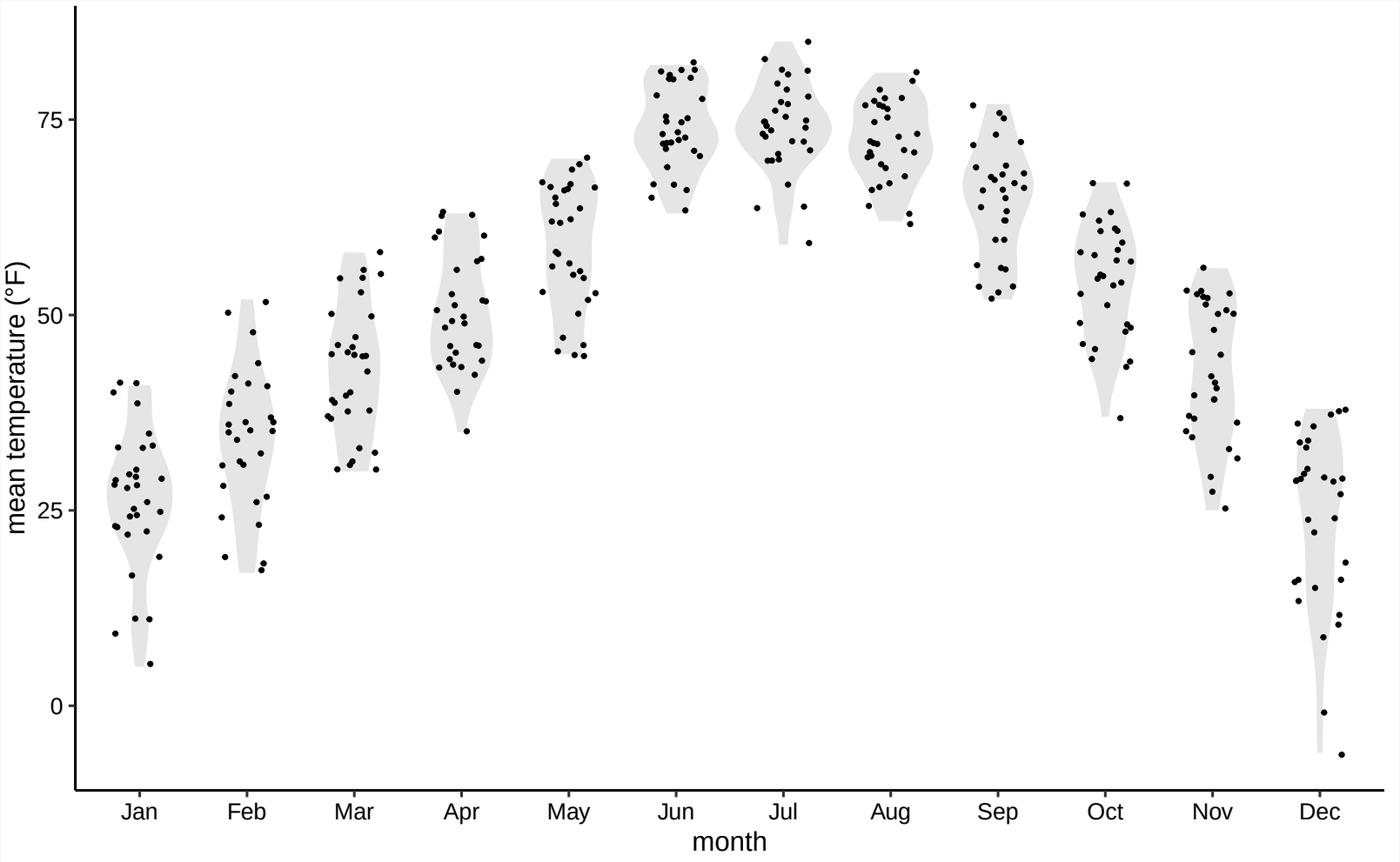
小提琴图


75
50
25
0
mean temperature (°F)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month

抖散图


mean temperature (°F)
75
50
25
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month

sina 图


mean temperature (°F)
75
50
25
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month

山峦图

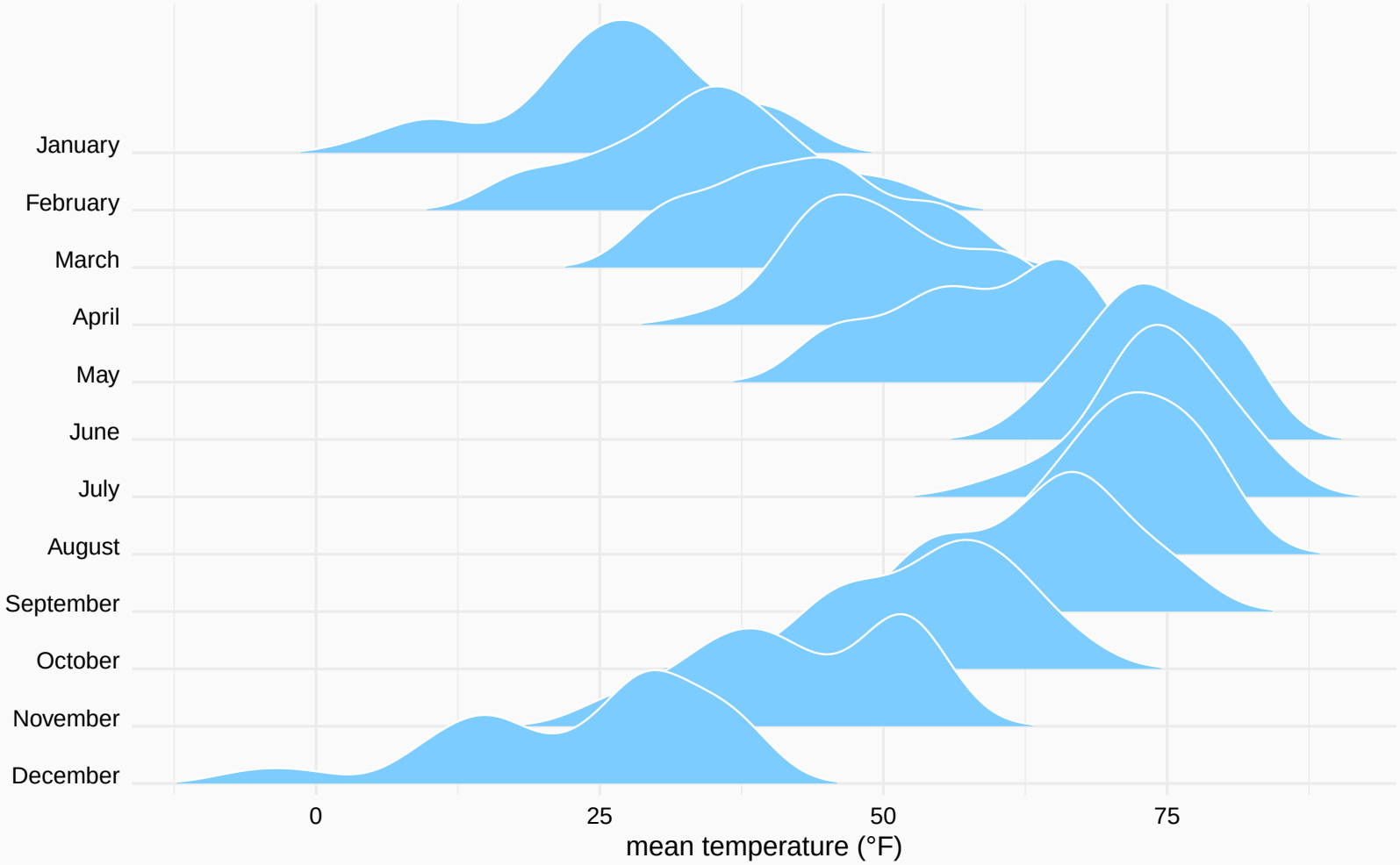


图 3: 林肯市 2016 年气温分布山峦图

有颜色山峦图

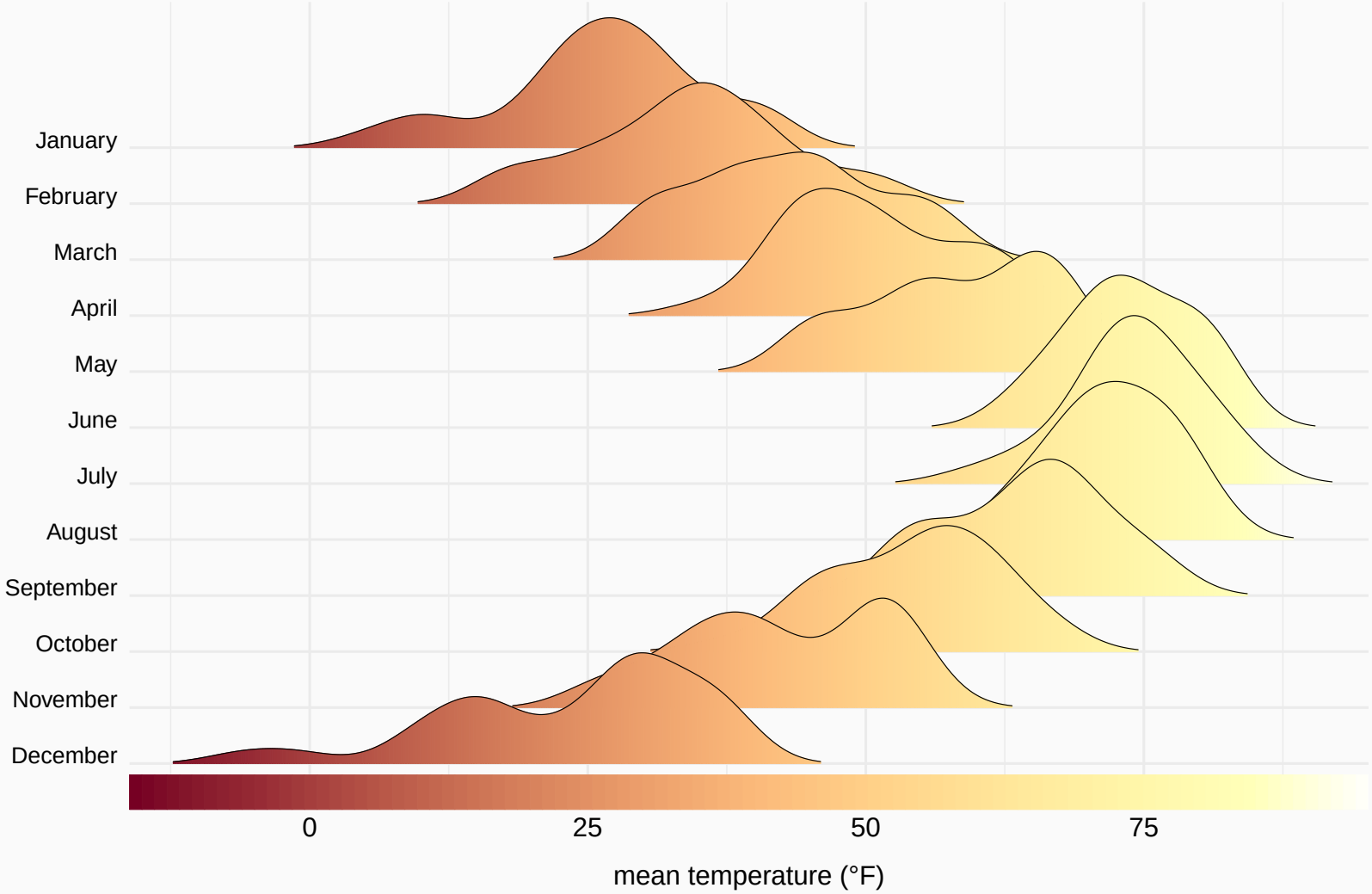
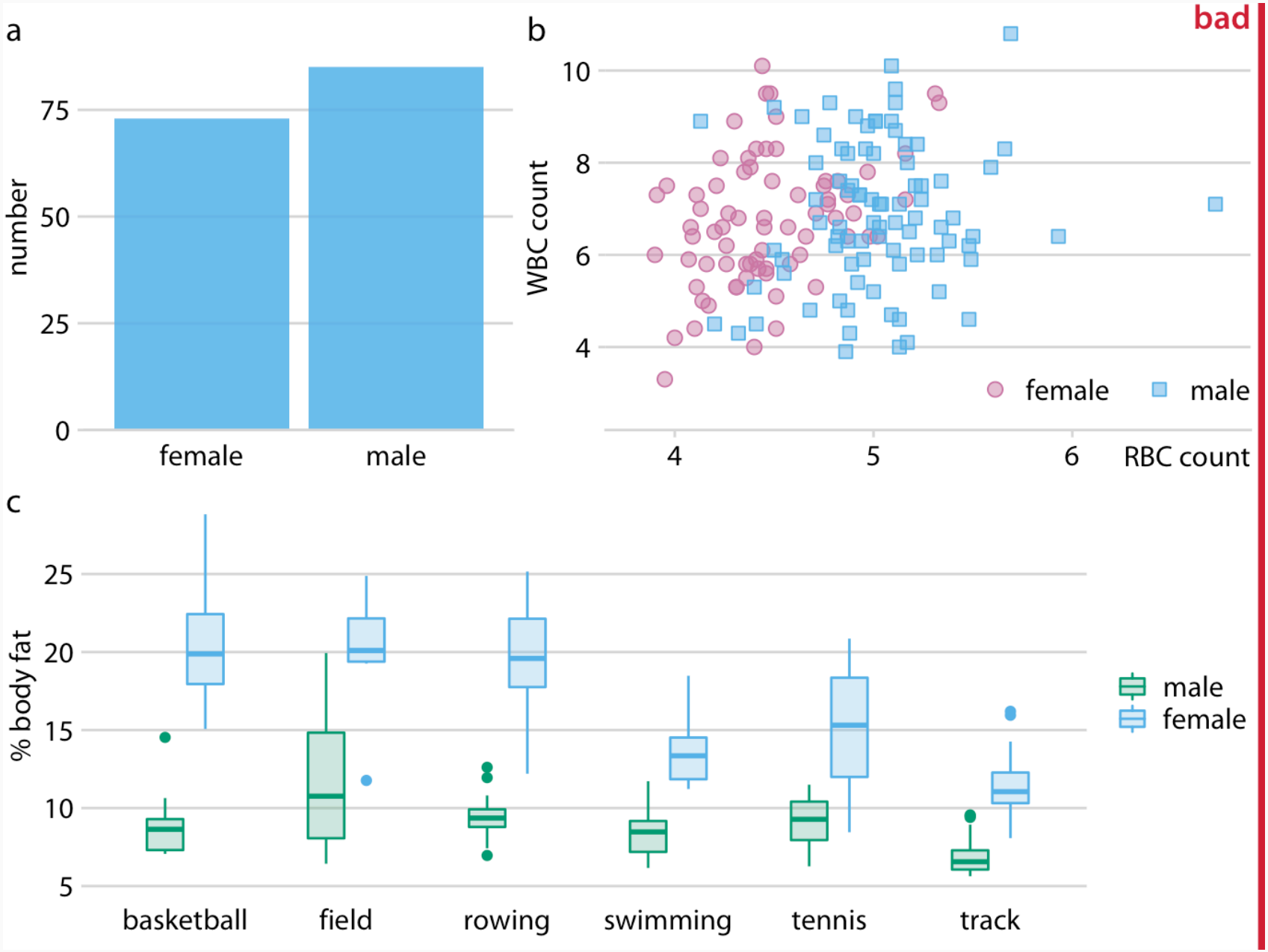
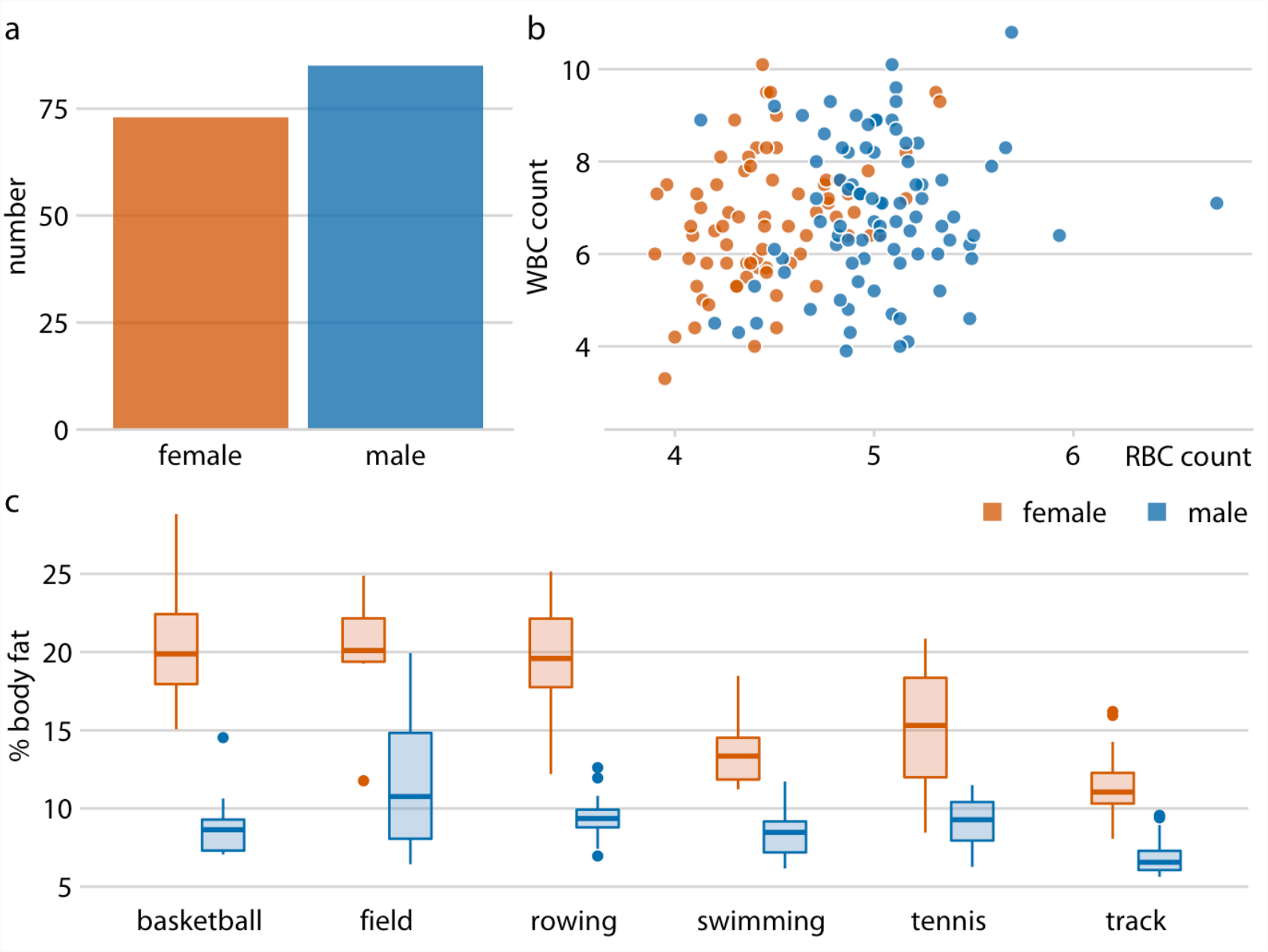


图 4: 林肯市 2016 年气温分布山峦图（颜色越亮，温度越高）

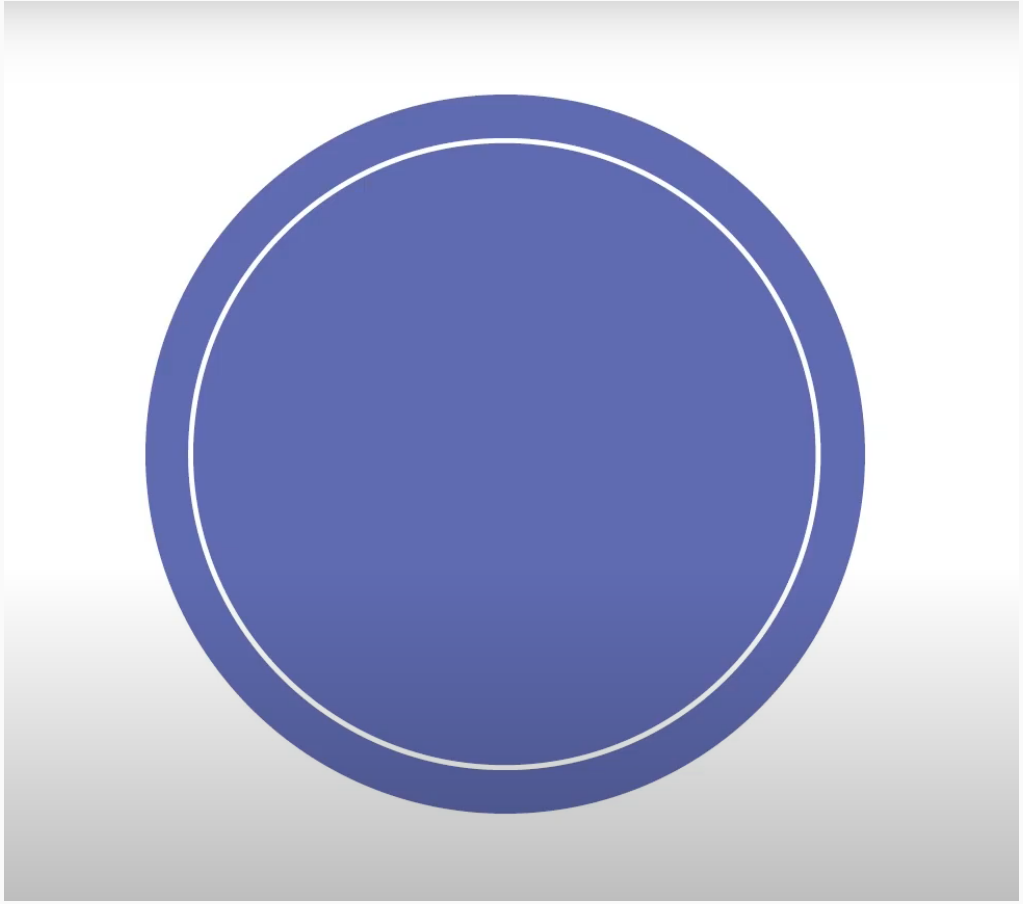
图片组合


a
number
75
50
25
0
female
male
b
bad
WBC count
10
8
6
4
4
5
6
RBC count
female
male
c
% body fat
25
20
15
10
5
basketball
field
rowing
swimming
tennis
track
male
female

图片组合

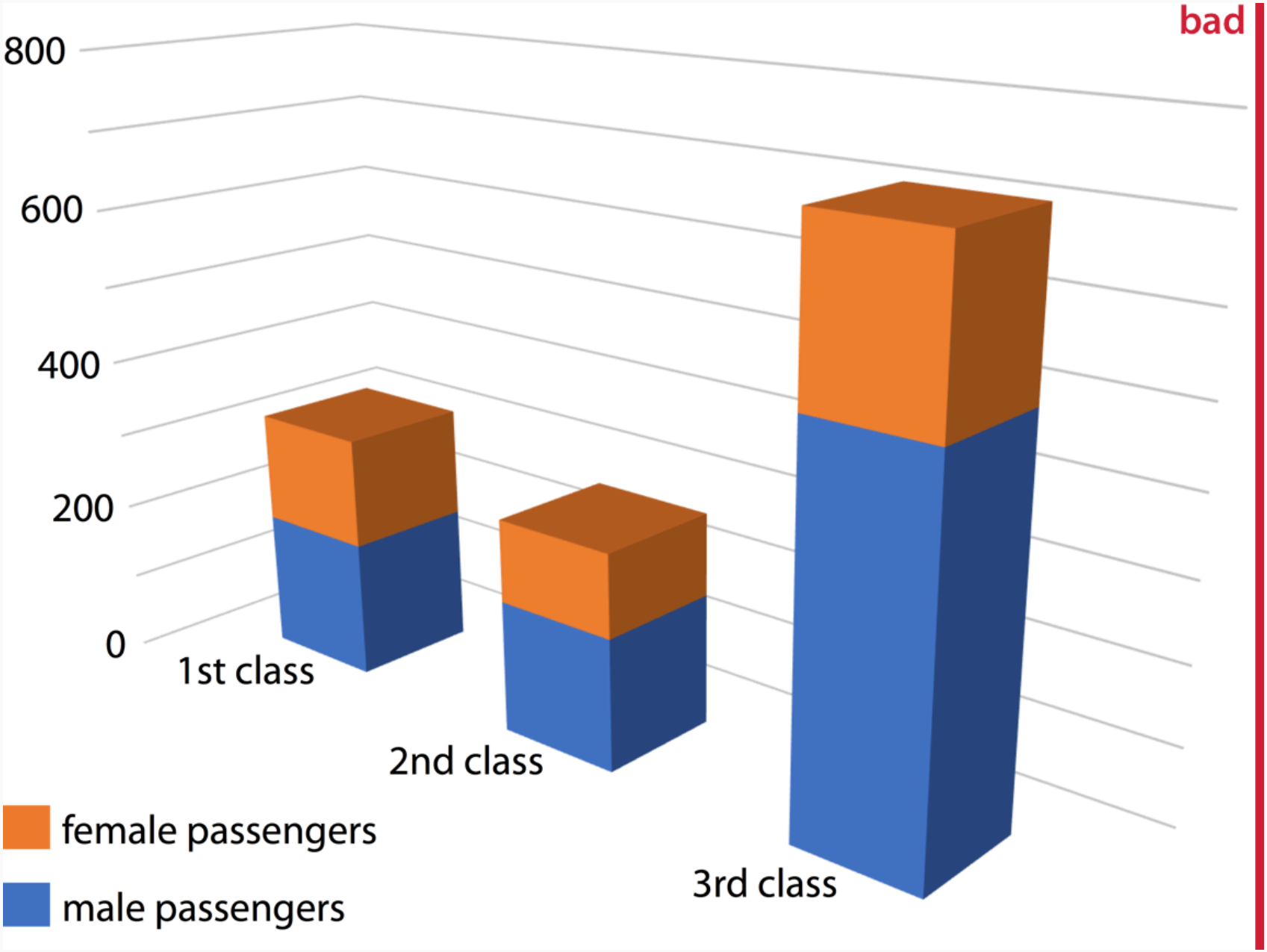

a
number
75
50
25
0
female
male
b
WBC count
10
8
6
4
4
5
6
RBC count
c
female
male
% body fat
25
20
15
10
5
basketball
field
rowing
swimming
tennis
track

尽可能不用 3D 图

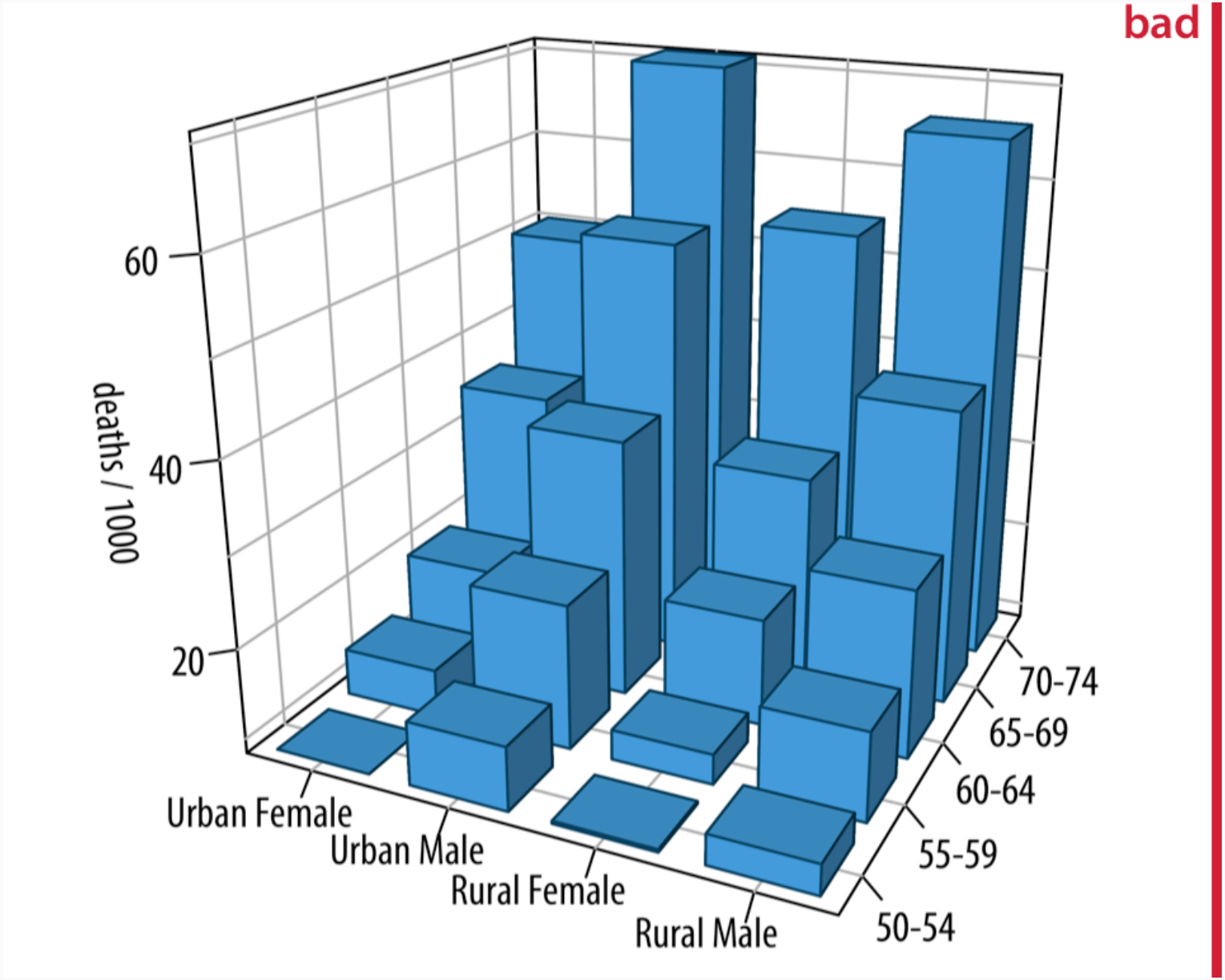


80% - 90% ?

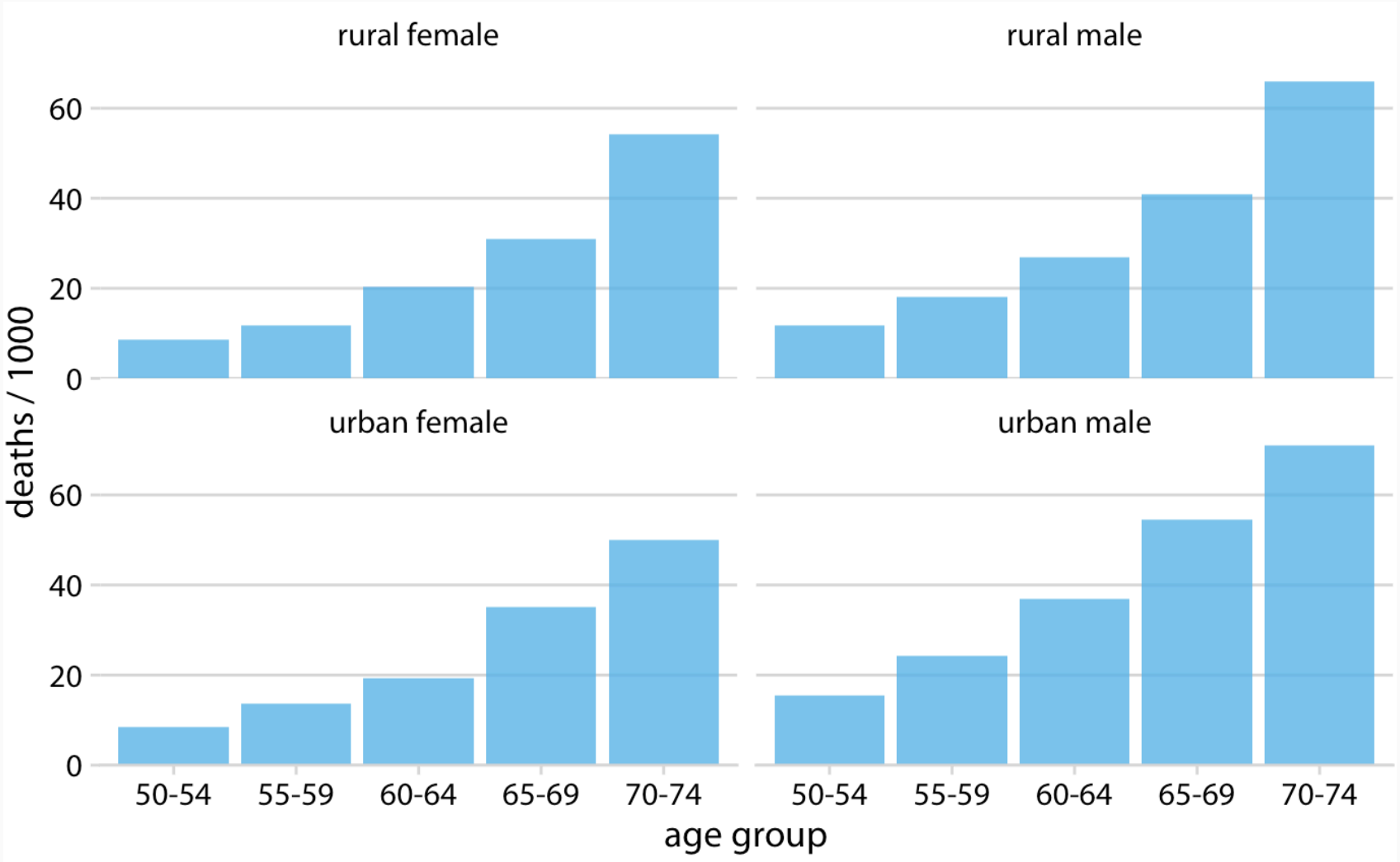
尽可能不用 3D 图


bad
800
600
400
200
0
1st class
2nd class
3rd class
female passengers
male passengers

尽可能不用 3D 图


bad
60
40
20
deaths / 1000
Urban Female
Urban Male
Rural Female
Rural Male
70-74
65-69
60-64
55-59
50-54

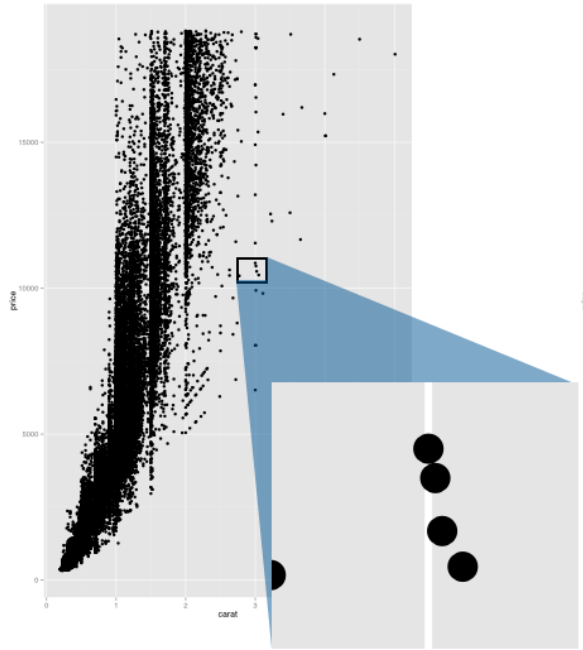
尽可能不用 3D 图


rural female
rural male
urban female
urban male
deaths / 1000
0
20
40
60
50-54
55-59
60-64
65-69
70-74
age group

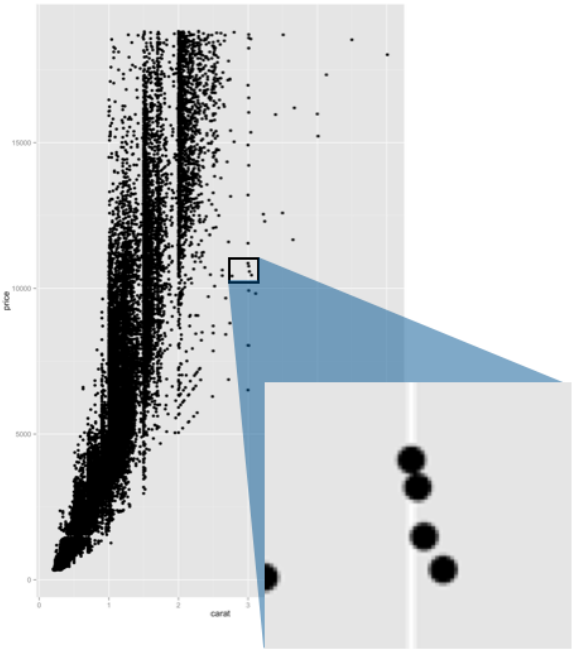
图片格式

Acronym	Name	Type	Application
pdf	Portable Document Format	vector	general purpose
eps	Encapsulated PostScript	vector	general purpose, outdated; use pdf
svg	Scalable Vector Graphics	vector	online use
png	Portable Network Graphics	bitmap	optimized for line drawings
jpeg	Joint Photographic Experts Group	bitmap	optimized for photographic images
tiff	Tagged Image File Format	bitmap	print production, accurate color reproduction
raw	Raw Image File	bitmap	digital photography, needs post-processing
gif	Graphics Interchange Format	bitmap	outdated for static figures, Ok for animations

图片格式


15000
10000
5000
0
price
0
1
2
3
carat

pdf


15000
10000
5000
0
price
0
1
2
3
carat

png

小节

准确传递信息，同时不增加读者心智负担

- 比例对等原则
- 排序很关键
- 处理好重叠点
- 色彩的运用
- 必要的标注
- 分布图的优缺点
- 图片组合
- 规避 3D 图
- 图片格式

代码实现

常用工具

工具	用途	使用
Excel, PowerBI, Tableau	商用	无需编程
Origin, Sigmaplot, GraphPad	学术用	无需编程
R, Python, matlab	编程工具	编程
Echarts, G2, D3.js	网页交互	编程

R 是什么

R 语言是用于统计分析，图形表示和报告的编程语言:

- R 是一个统计编程语言（statistical programming）
- R 可运行于多种平台之上，包括 Windows、UNIX 和 Mac OS X
- R 拥有顶尖水准的制图功能
- R 是免费的
- R 应用广泛，拥有丰富的库包
- 活跃的社区

官网定义：https://www.r-project.org/

R 路上的大神

2019 年 8 月，国际统计学年会将考普斯总统奖（被誉为统计学的诺贝尔奖）奖颁给 ggplot2 的作者

- Hadley Wickham，一个改变了 R 语言的人

ggplot2

ggplot2 是最受欢迎的 R 宏包，没有之一

```
library(cranlogs)
  d <- cran_downloads(
         package = "ggplot2",
            from = "2020-01-01",
              to = "2020-09-01"
        )

sum(d$count)
#> [1] 11540486
```

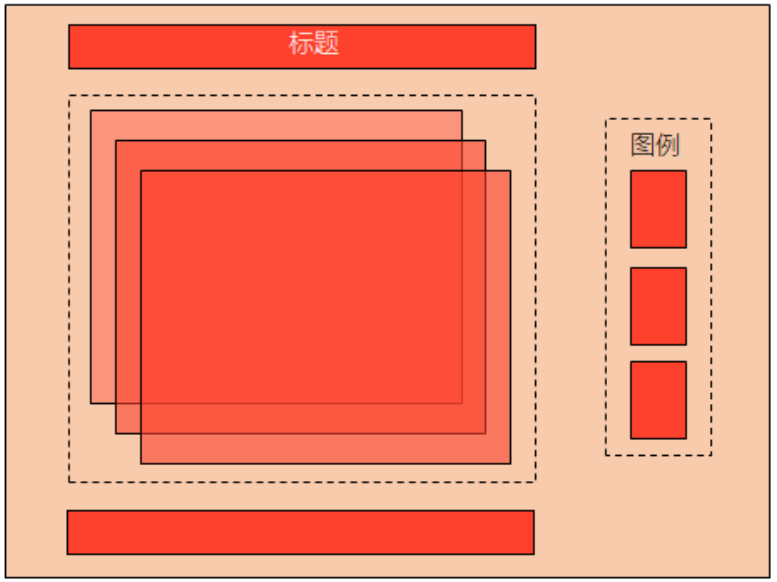
ggplot2 语法

绘制一幅水彩画

画布

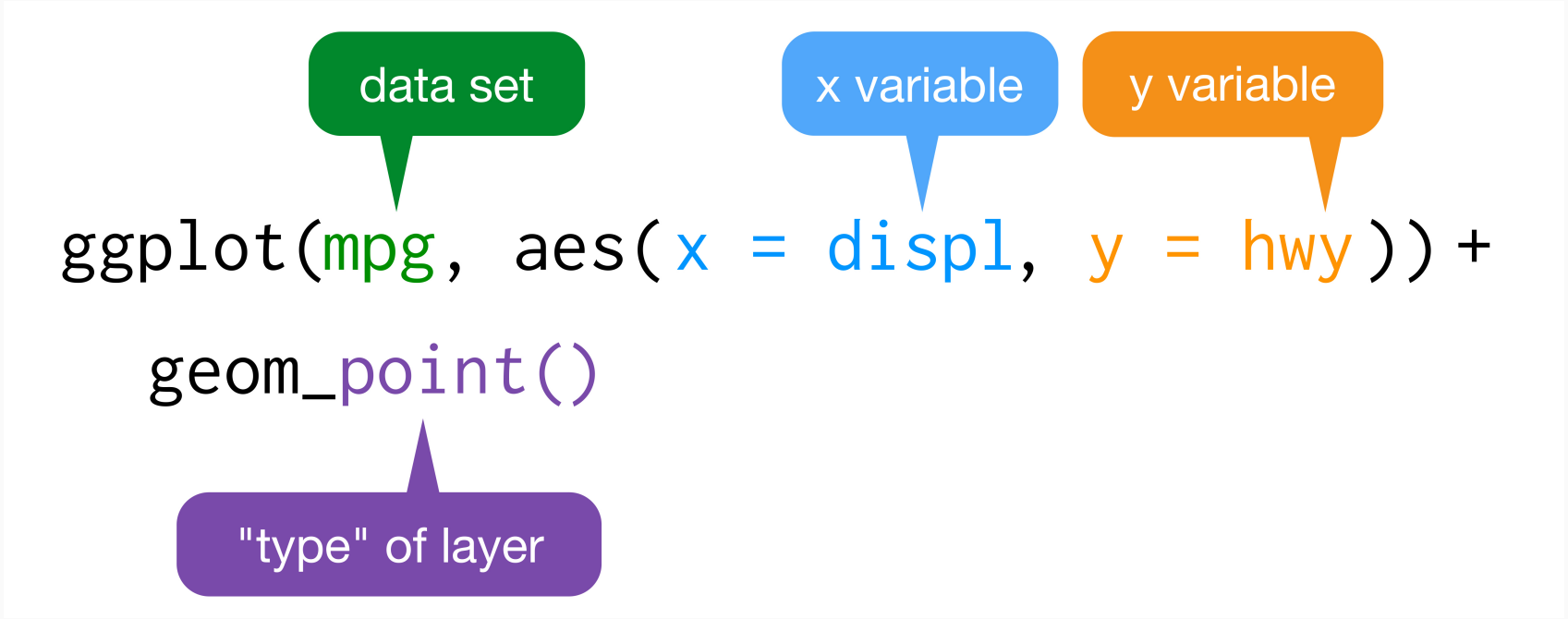
- 在画板上摊开画布
- 使用笔刷涂刷颜料（第一层、第二层、第三层、.....第N层）
- 修饰局部与美化细节

绘制一幅图表



- 初始化图层——ggplot()
- 添加特定几何图层geom_XXX（散点图、条形图、面积图等）
- 修饰局部与美化细节(优化配色、字体、布局和图形元素)

ggplot2 语法

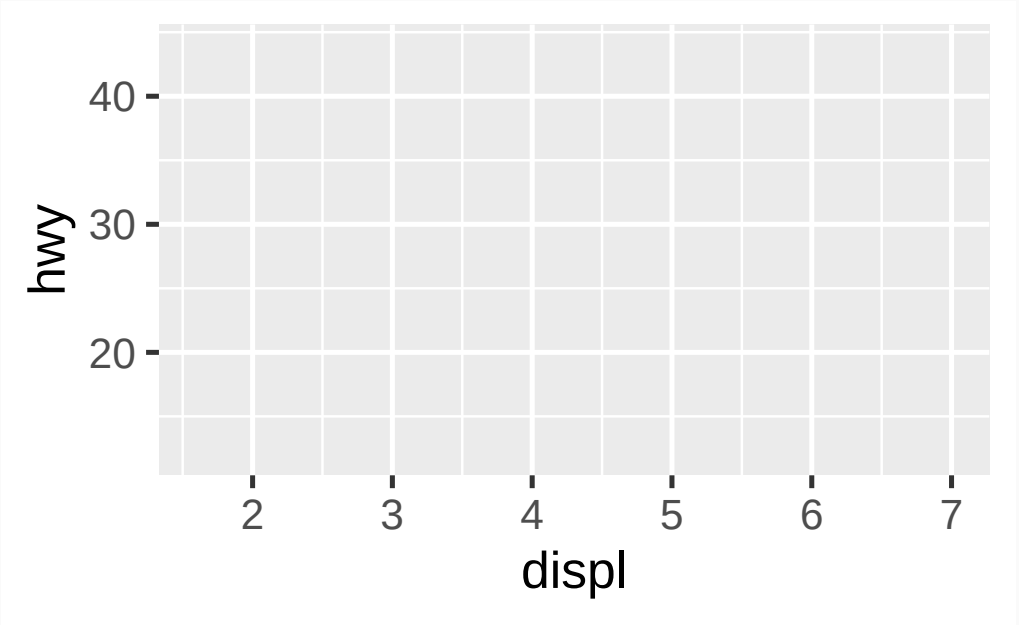

data set
x variable
y variable
ggplot(mpg, aes(x = displ, y = hwy)) +
geom_point()
"type" of layer

ggplot2 语法很容易理解

```
ggplot()
```

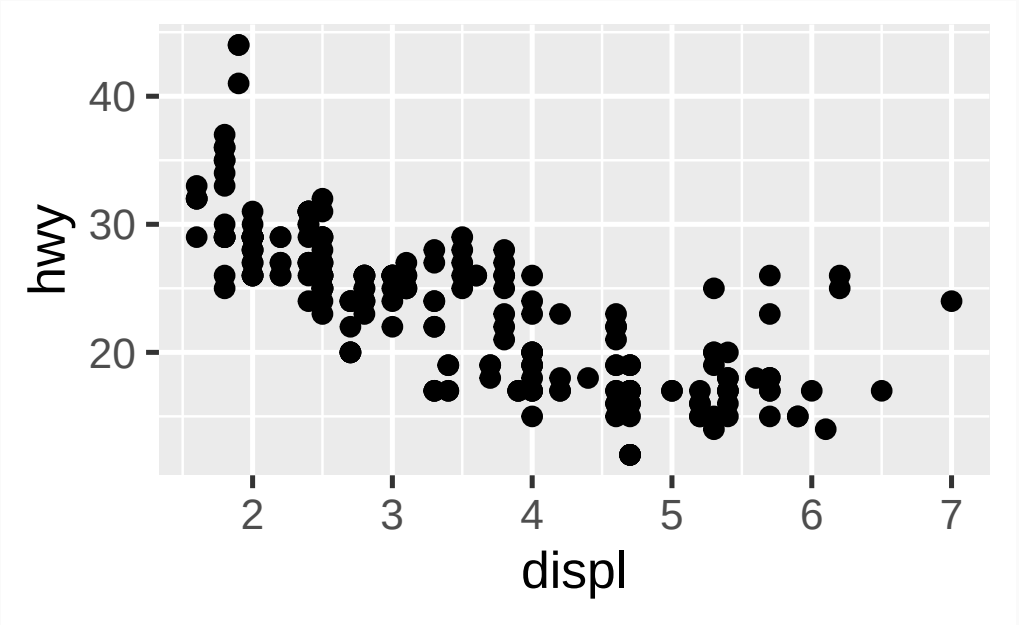
ggplot2 语法很容易理解

```
ggplot(mpg, aes(displ, hwy))
```



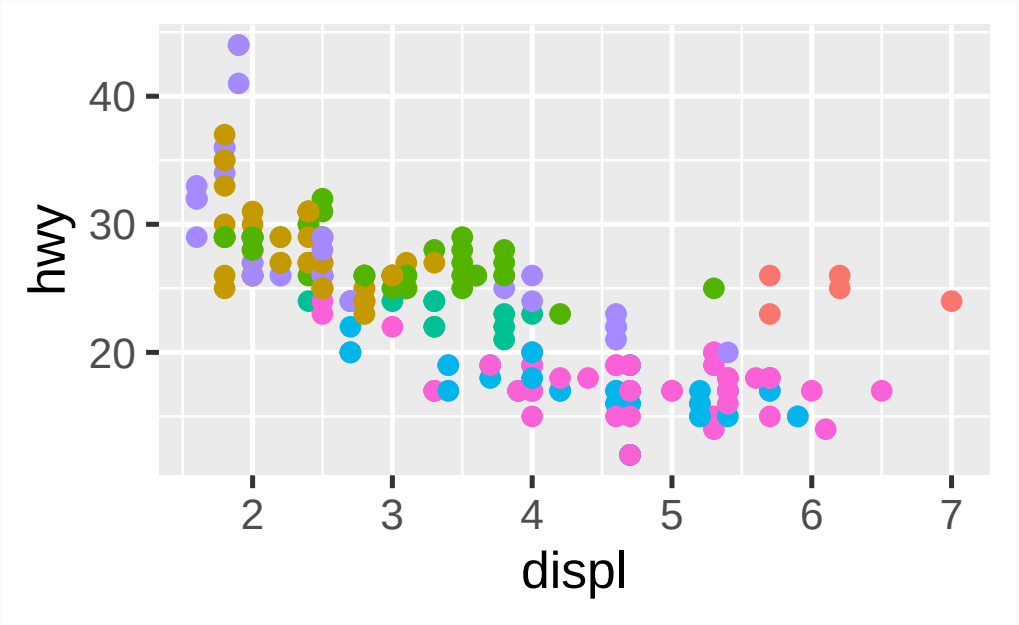
ggplot2 语法很容易理解

```
ggplot(mpg, aes(displ, hwy)) +
  geom_point()
```



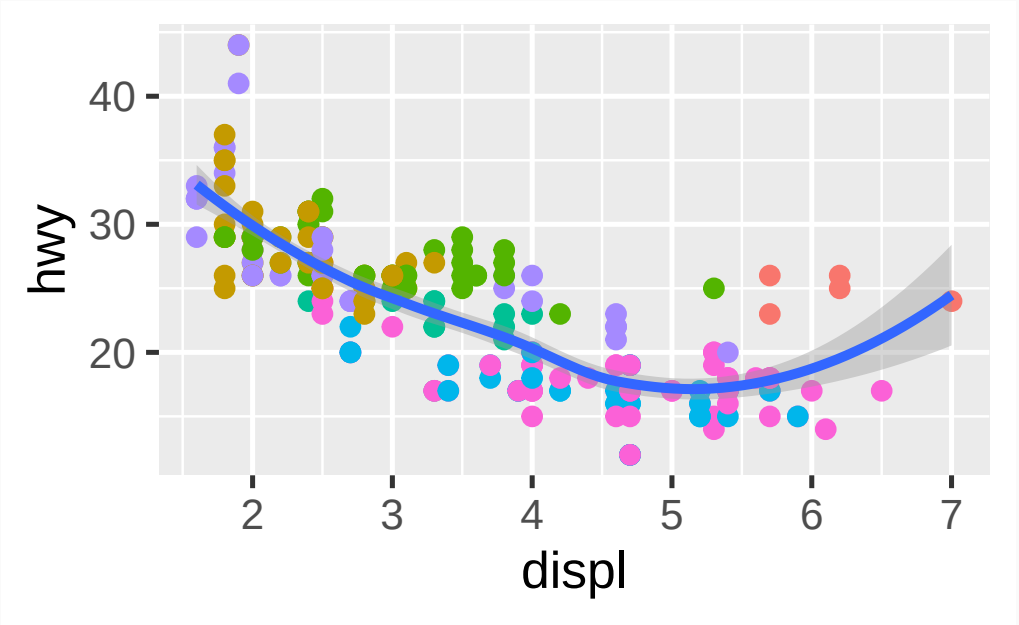
ggplot2 语法很容易理解

```
ggplot(mpg, aes(displ, hwy)) +
  geom_point(aes(color = class))
```



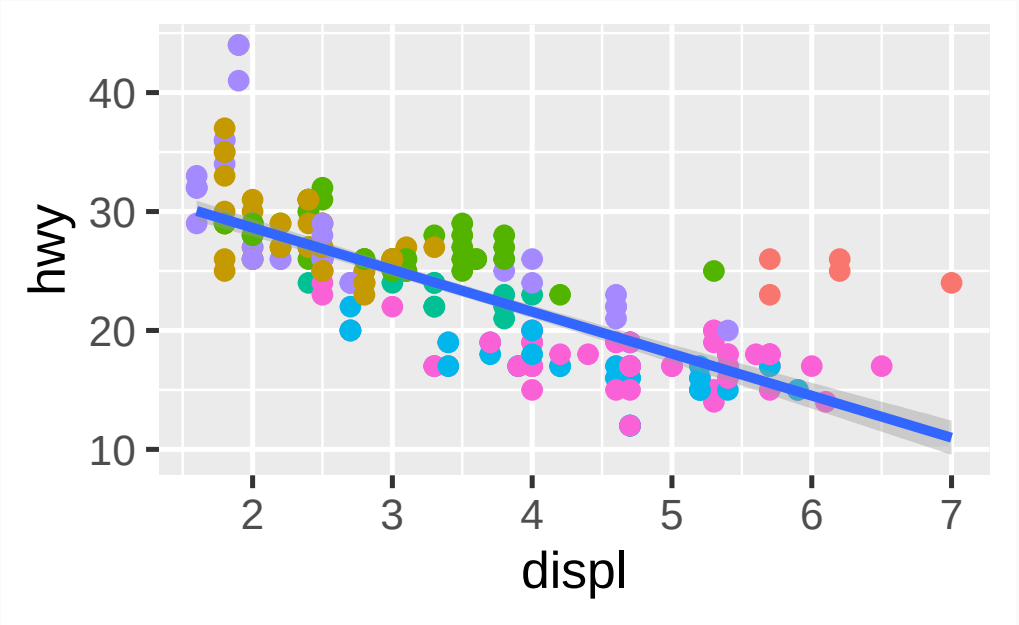
ggplot2 语法很容易理解

```
ggplot(mpg, aes(displ, hwy)) +
  geom_point(aes(color = class)) +
  geom_smooth()
```



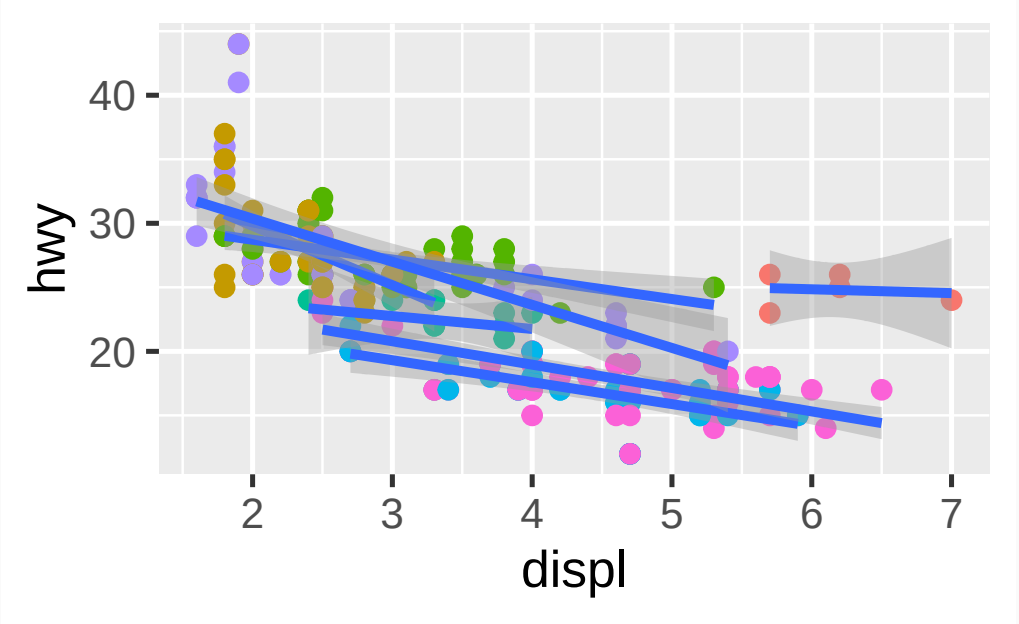
ggplot2 语法很容易理解

```
ggplot(mpg, aes(displ, hwy)) +
  geom_point(aes(color = class)) +
  geom_smooth(
    method = "lm"
  )
```



ggplot2 语法很容易理解

```
ggplot(mpg, aes(displ, hwy)) +
  geom_point(aes(color = class)) +
  geom_smooth(
    method = "lm",
    aes(group = class)
  )
```



一见钟情，还是相见恨晚？

推荐书目

强烈推荐

- R for Data Science
- ggplot2: Elegant Graphics for Data Analysis
- R Graphics Cookbook
- Fundamentals of Data Visualization
- Data Visualization: A practical introduction

感谢《Fundamentals of Data Visualization》作者 Claus O. Wilke，为大家写了这本非常好的书

join us

《数据科学中的 R 语言》讲解 R 语言入门基础、数据可视化、数据处理、探索性分析、统计建模以及在代表性领域的应用

- 课程性质：研究生公选课
- 上课地点：狮子山校区/成龙校区
- 课程内容：https://bookdown.org/wangminjie/R4DS/
- 学时学分：32/2
- 选课方式：研究生院

我会努力的

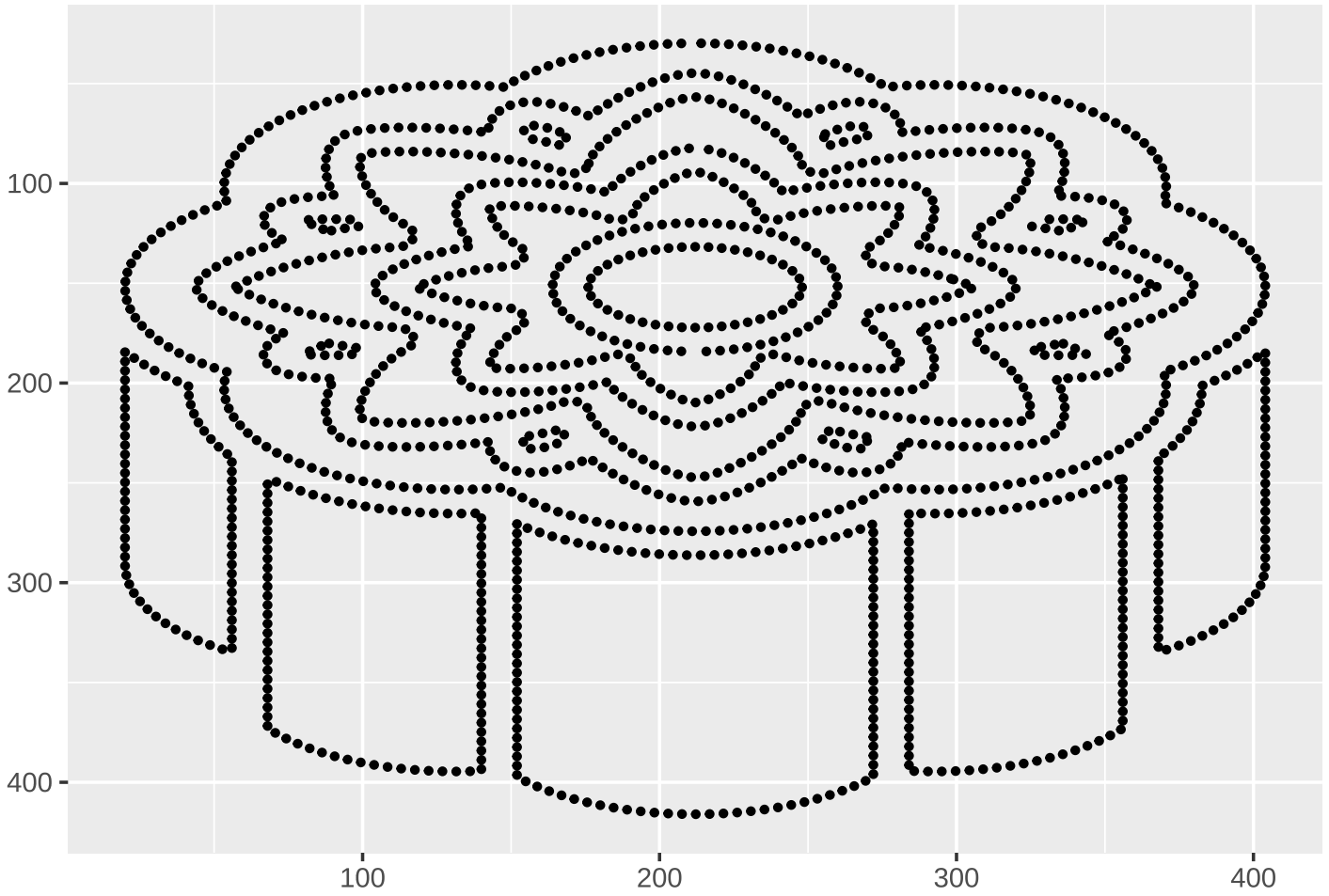
愿 R 语言成为你构建知识大厦的脚手架!

38552109@qq.com

谢谢观看

祝大家中秋节/国庆节快乐！

感谢四川师范大学研究生院和图书馆的信任和支持。



感谢 biobabble 宏包提供月饼数据