

R 语言科研数据可视化

王敏杰

2020 年 9 月 27 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

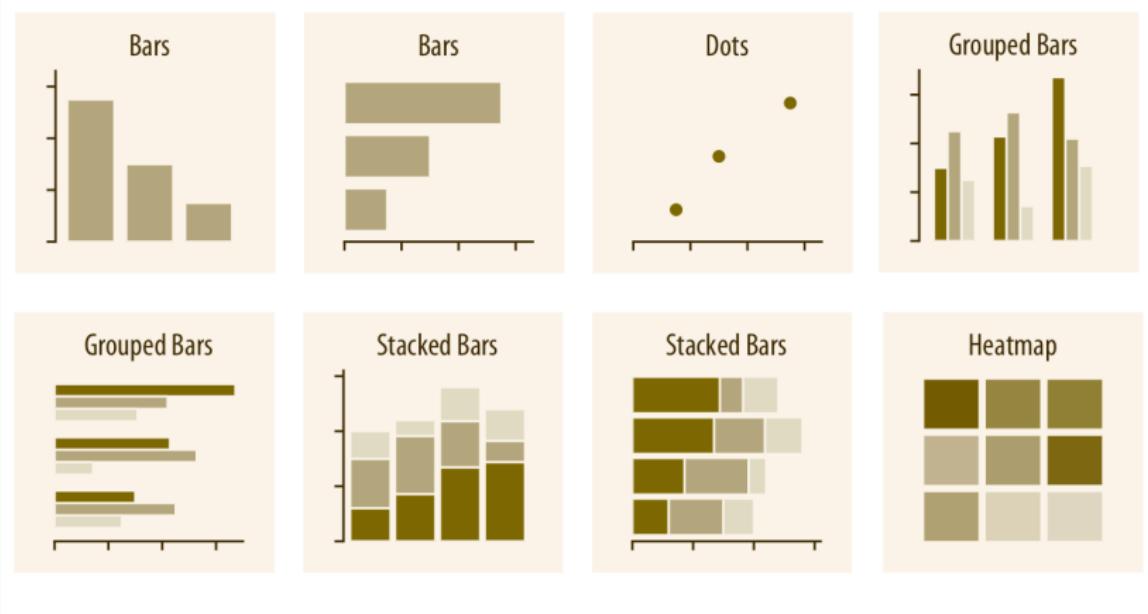
基本图形

基本图形

用什么样的图，取决于我们的需求

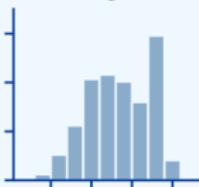
- 呈现数量
- 呈现分布
- 呈现比例
- 呈现关联
- 呈现不确定性
- 呈现时间变化
- 呈现地理信息

呈现数量

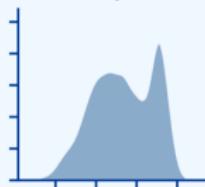


分布图

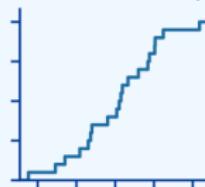
Histogram



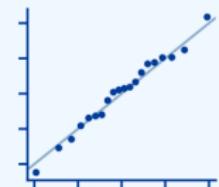
Density Plot



Cumulative Density

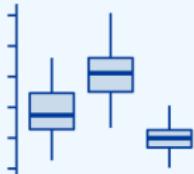


Quantile-Quantile Plot

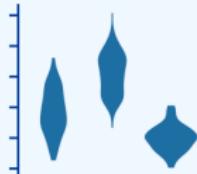


分布图

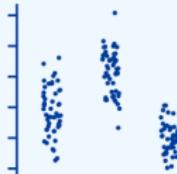
Boxplots



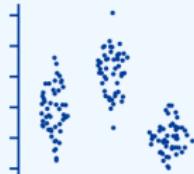
Violins



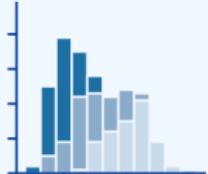
Strip Charts



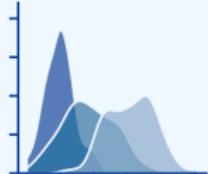
Sina Plots



Stacked Histograms



Overlapping Densities



Ridgeline Plot

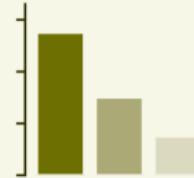


呈现比例

Pie Chart



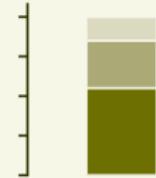
Bars



Bars



Stacked Bars

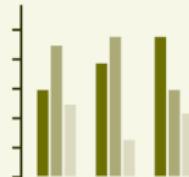


呈现比例

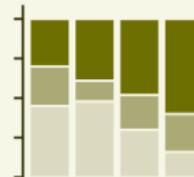
Multiple Pie Charts



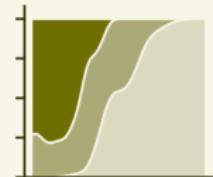
Grouped Bars



Stacked Bars



Stacked Densities

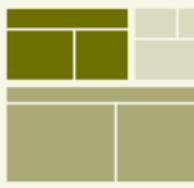


呈现比例

Mosaic Plot



Treemap



Parallel Sets

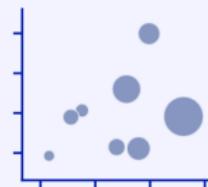


x-y 关联图

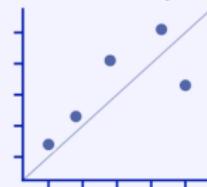
Scatterplot



Bubble Chart



Paired Scatterplot

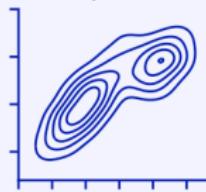


Slopegraph

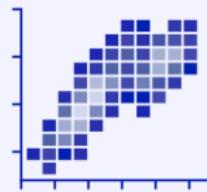


x-y 关联图

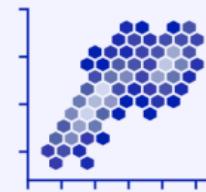
Density Contours



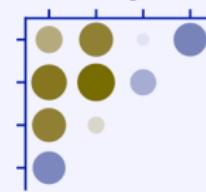
2D Bins



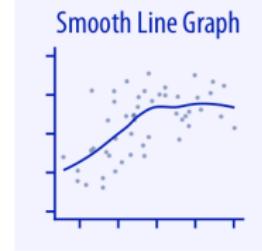
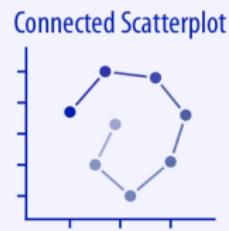
Hex Bins



Correlogram



x-y 关联图



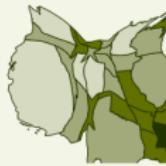
地图



Choropleth



Cartogram

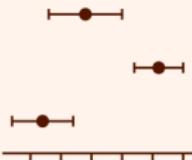


Cartogram Heatmap

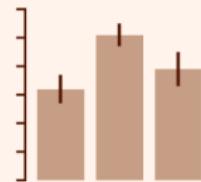


呈现不确定性

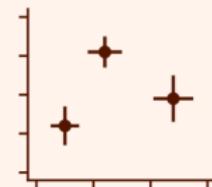
Error Bars



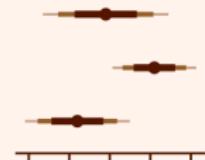
Error Bars



2D Error Bars



Graded Error Bars

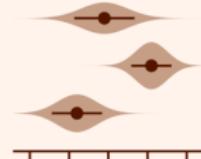


呈现不确定性

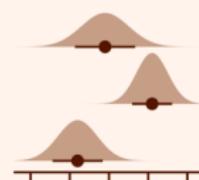
Confidence Strips



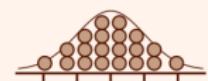
Eyes



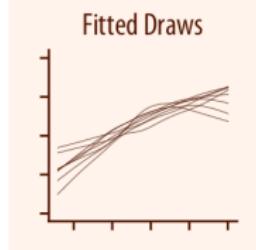
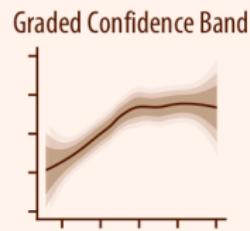
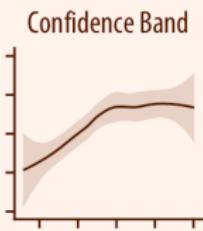
Half-Eyes



Quantile Dot Plot



呈现不确定性

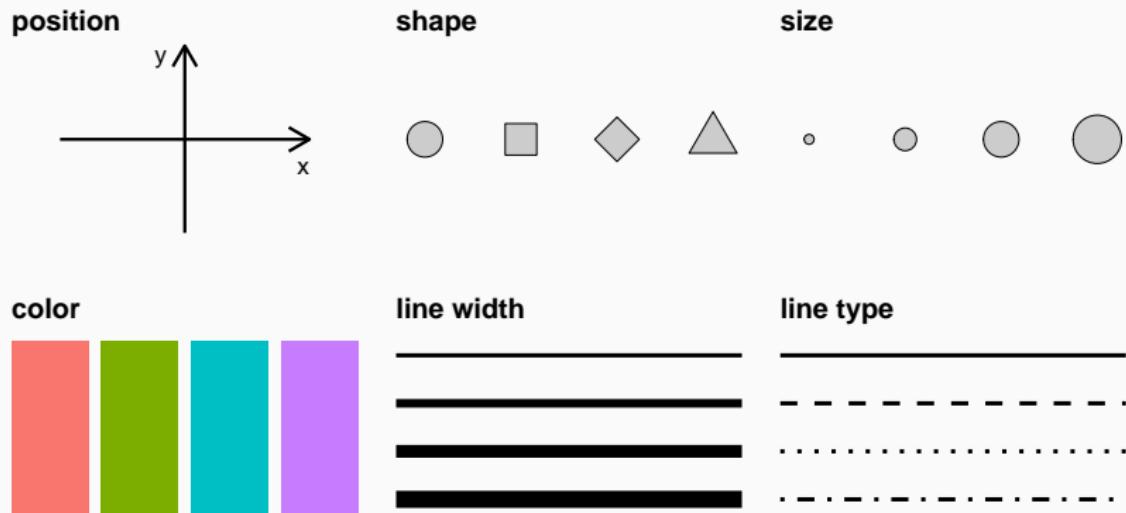


基本原则

基本原则

准确传递信息，同时不增加
读者心智负担

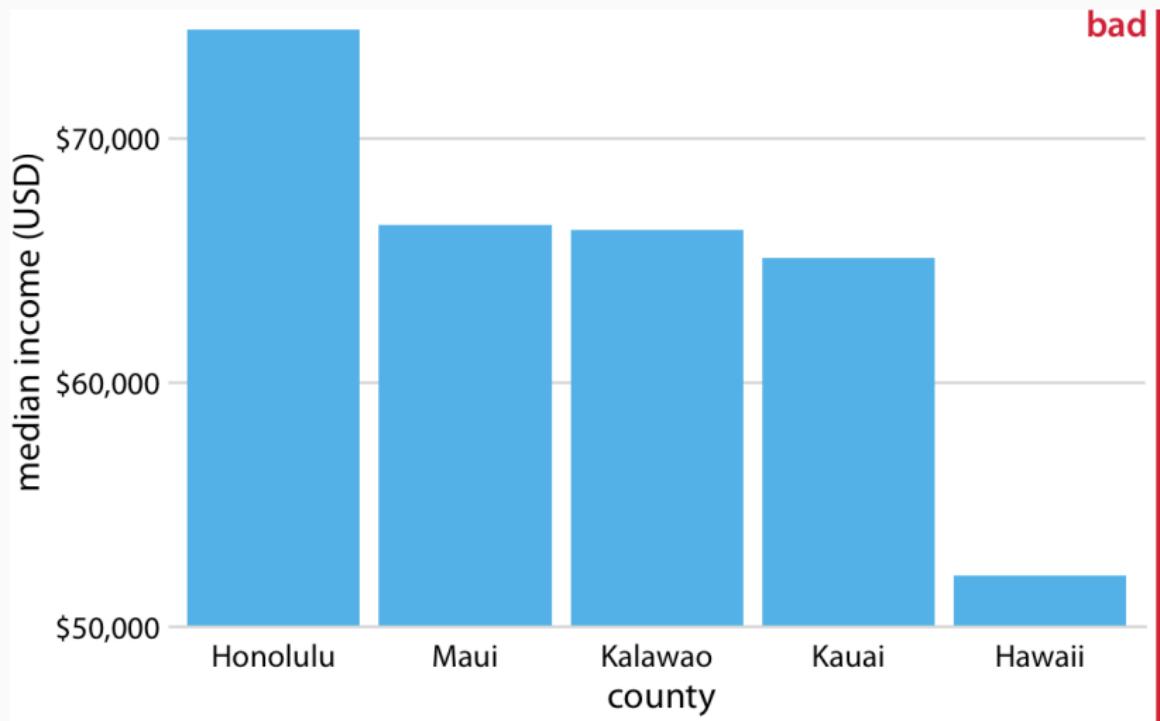
比例对等原则



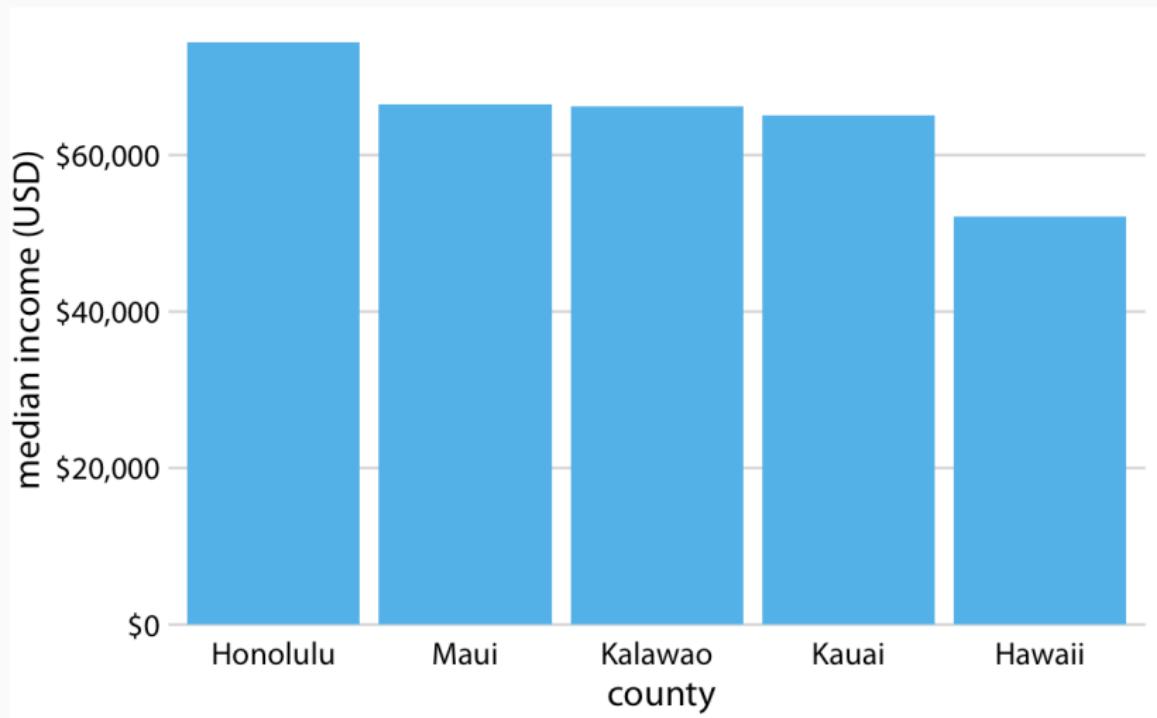
比例对等原则 (The principle of proportional ink):

- 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

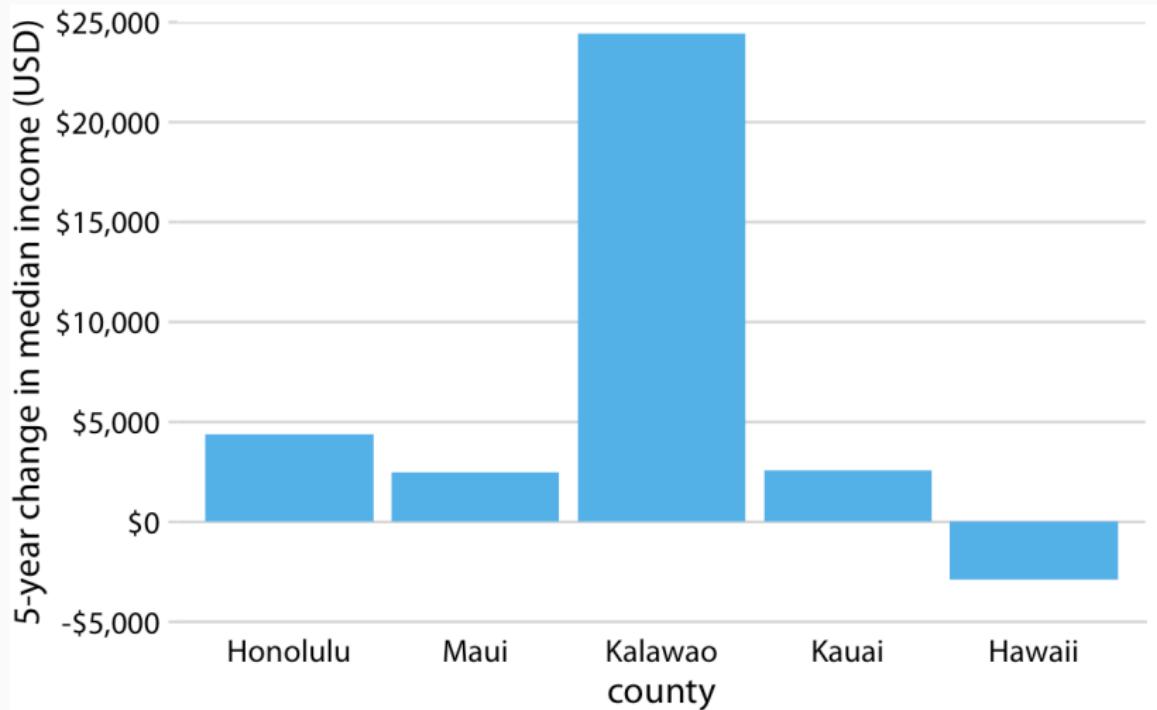


比例对等原则

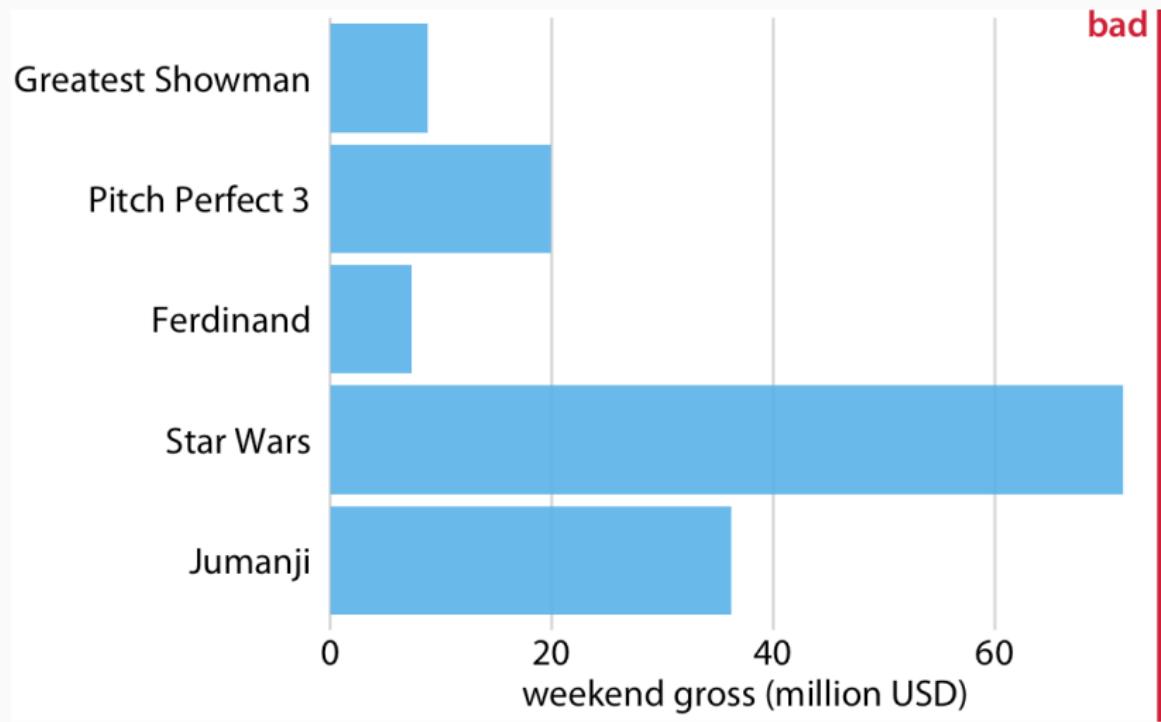


柱状图应该从 0 开始

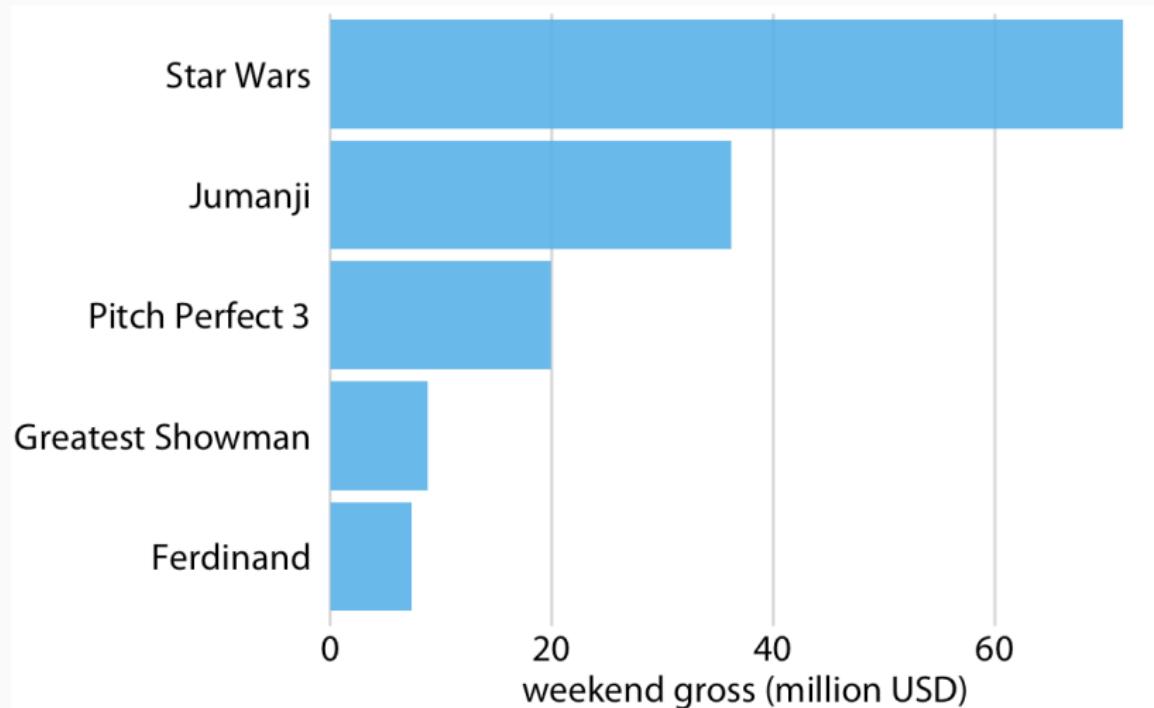
比例对等原则



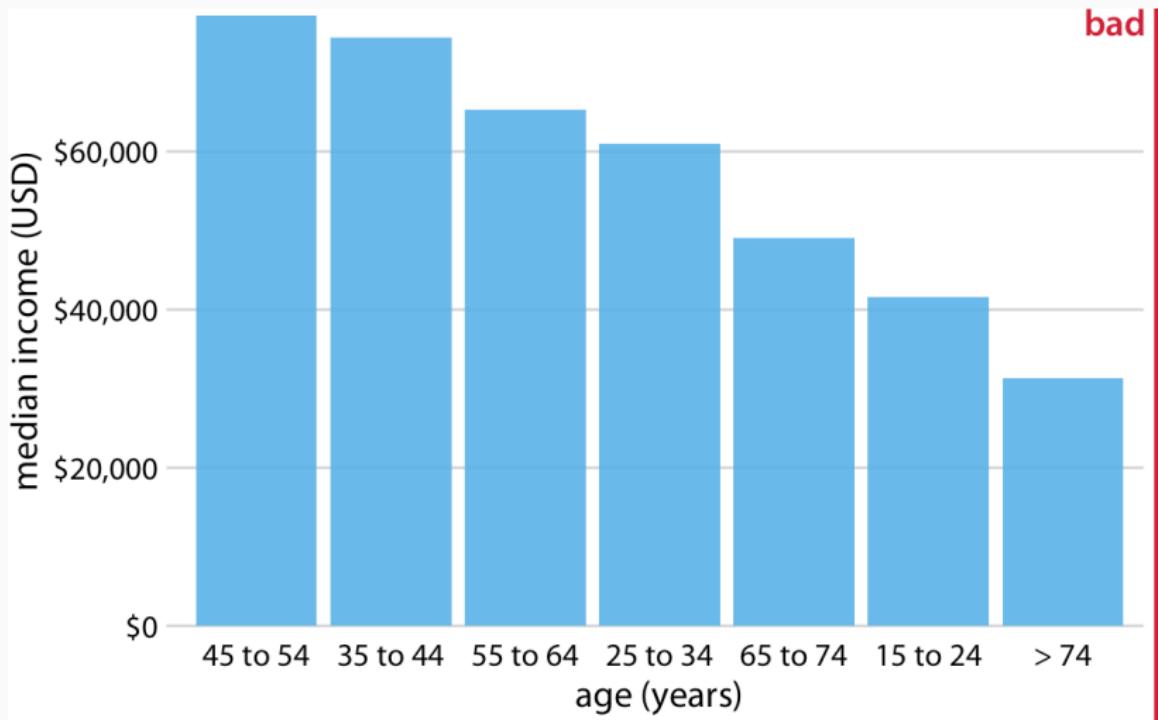
排序很关键



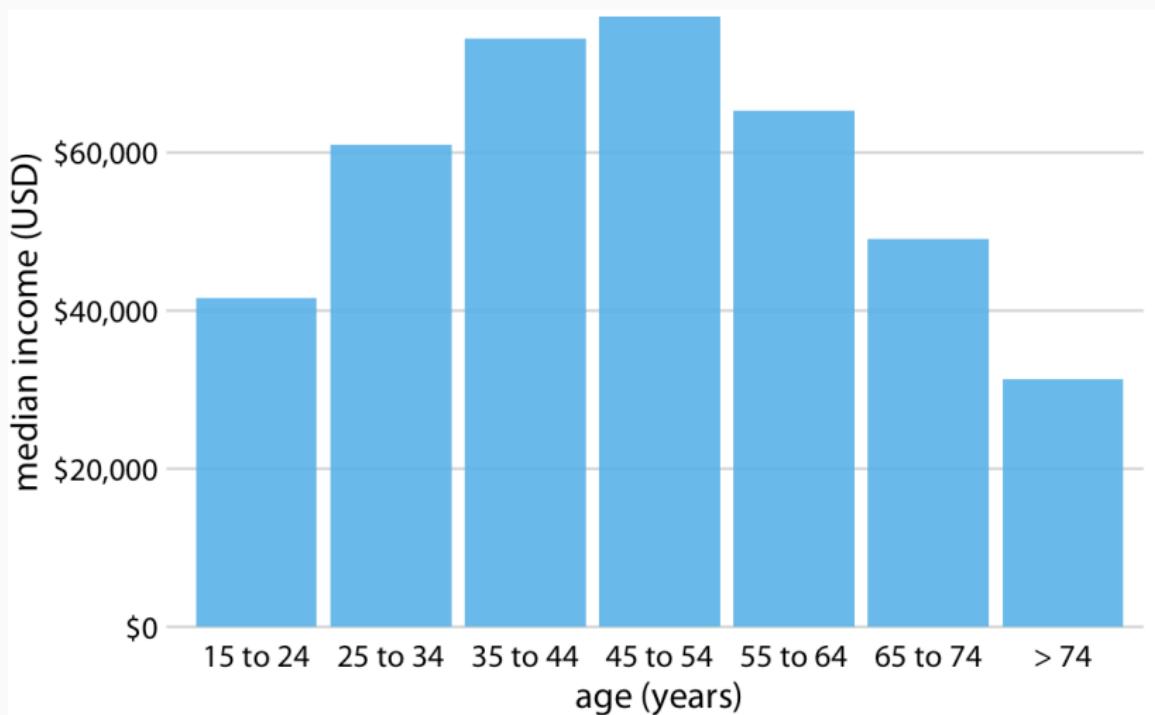
排序很关键



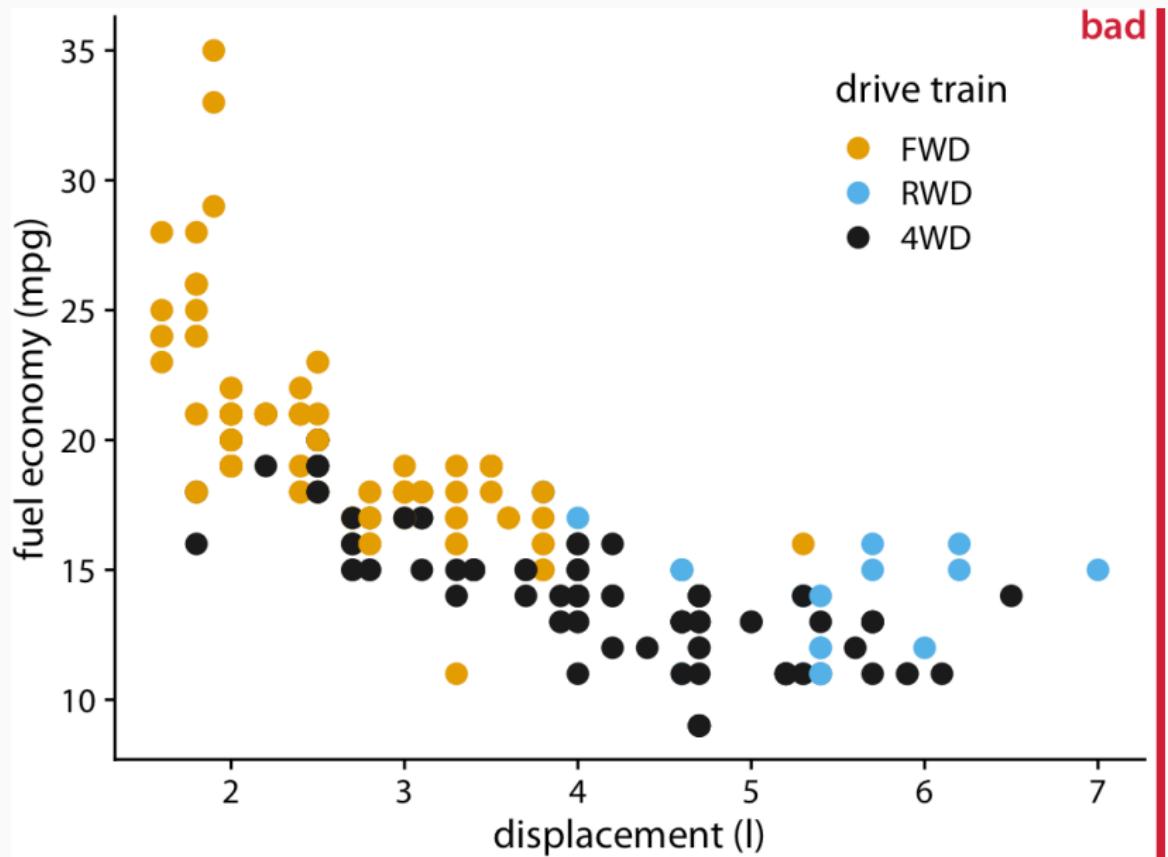
排序很关键



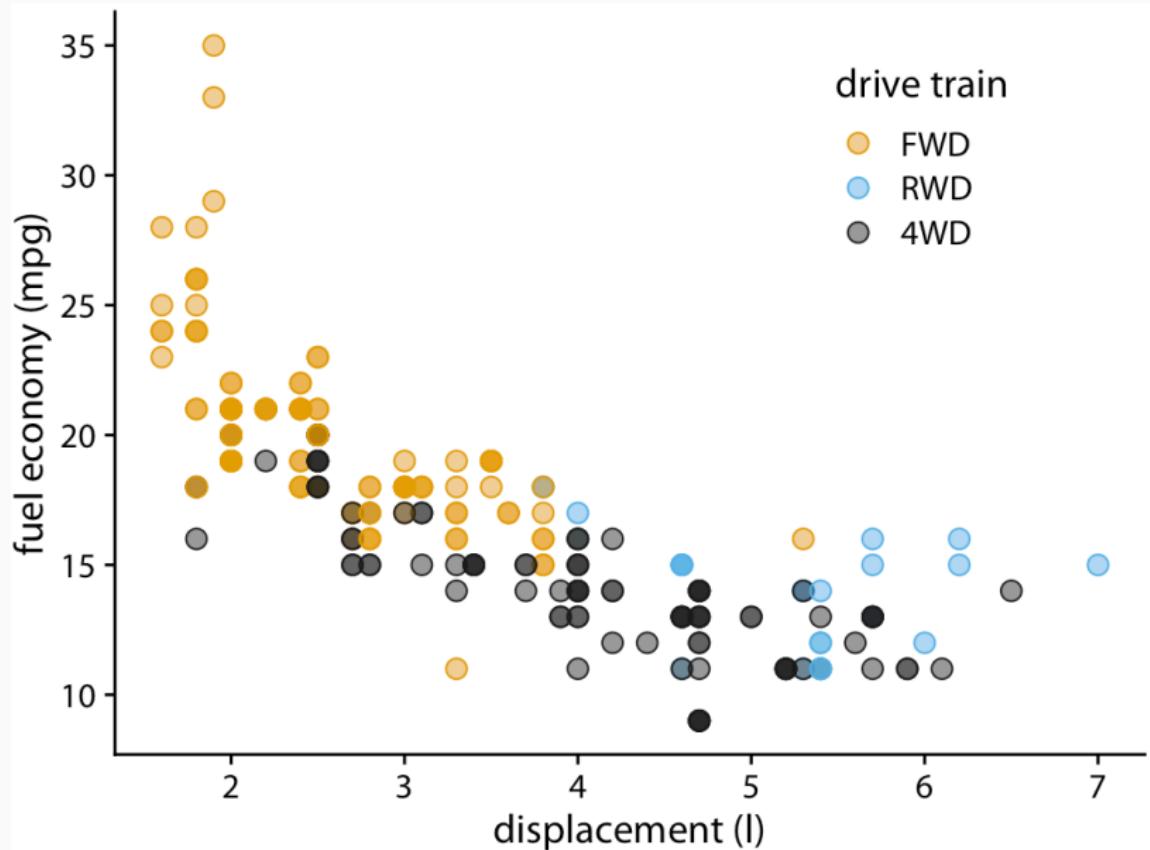
排序很关键



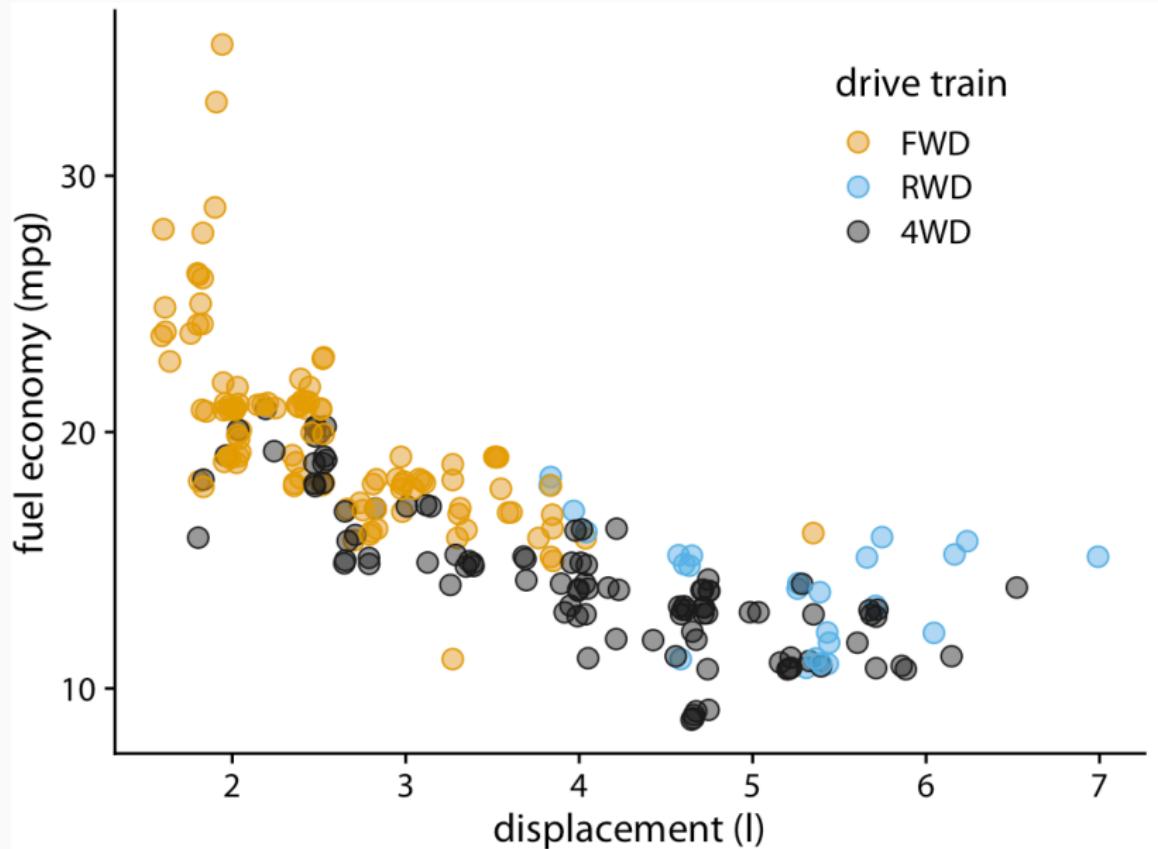
处理好重叠点



处理好重叠点

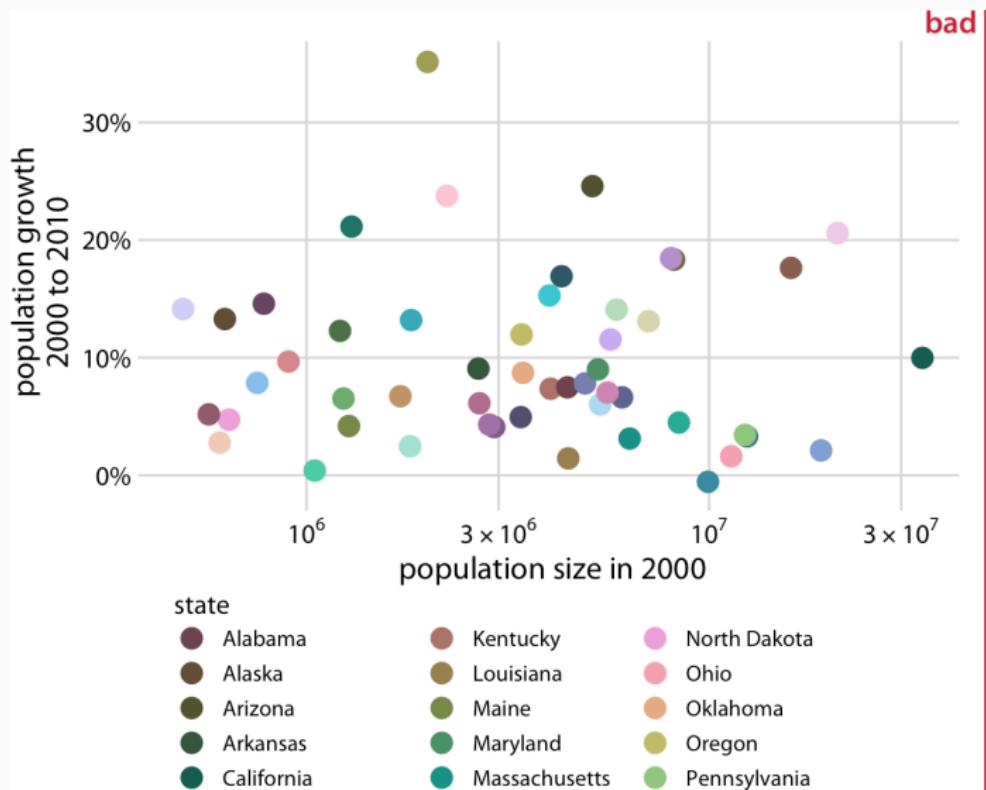


处理好重叠点



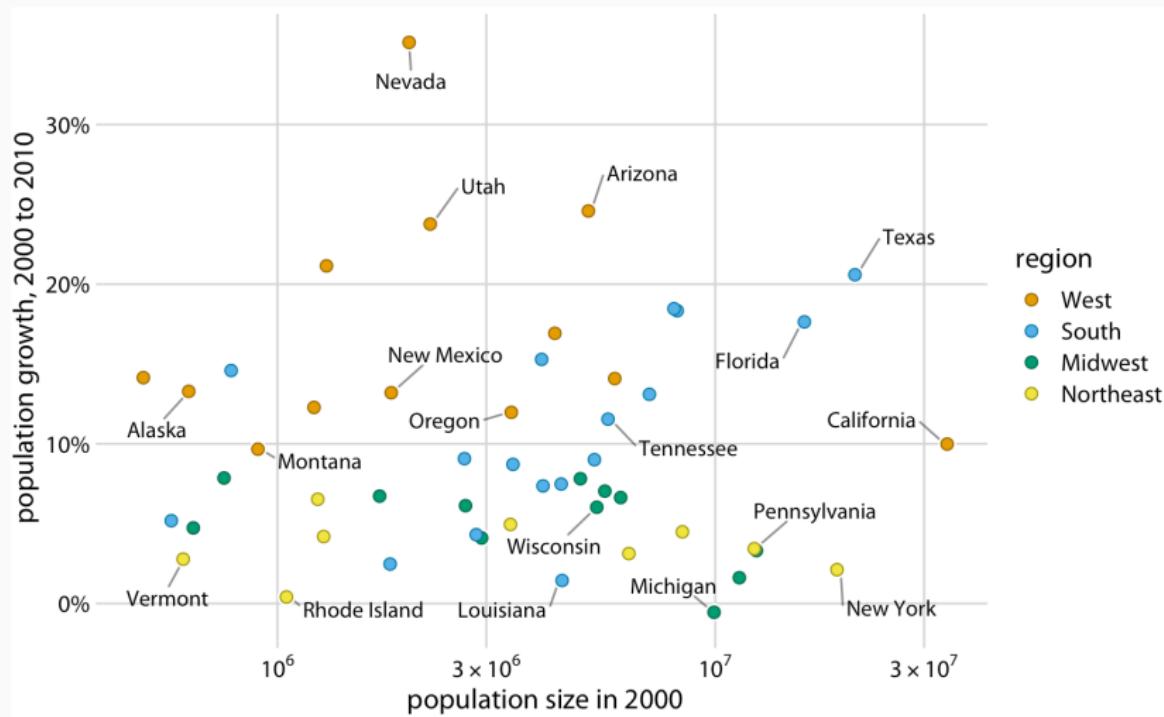
色彩的运用

颜色，可以提升可视化效果，也可毁掉本来很好的图

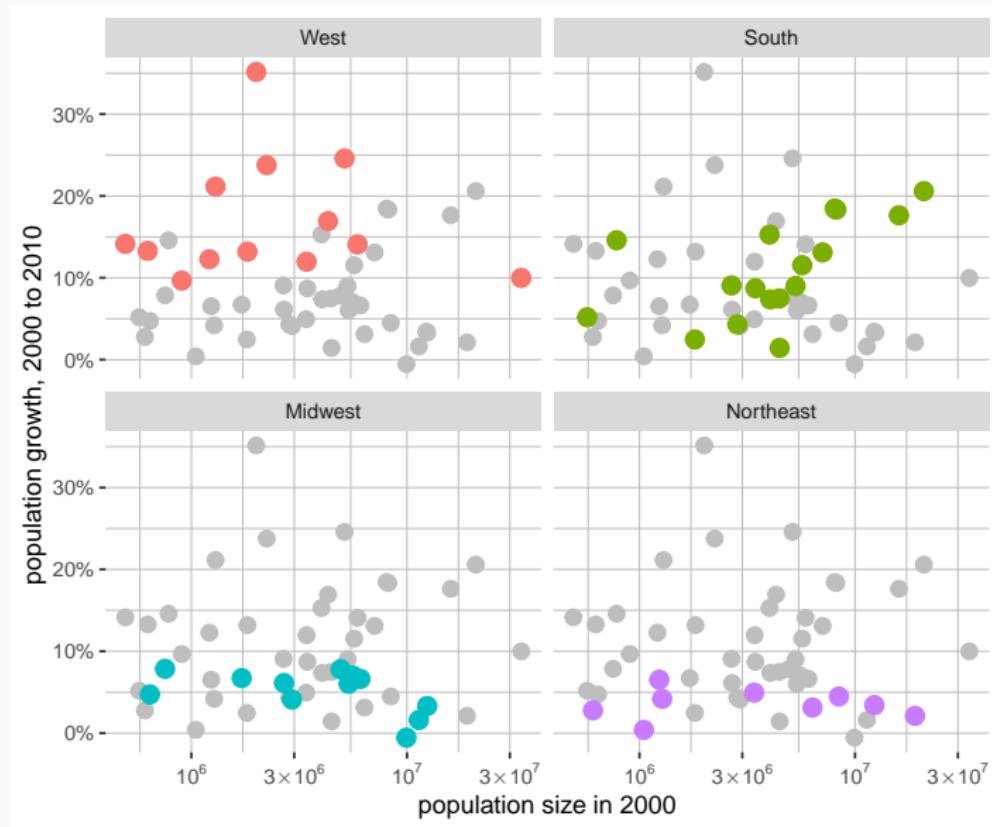


色彩的运用

颜色分组，同时增加标注

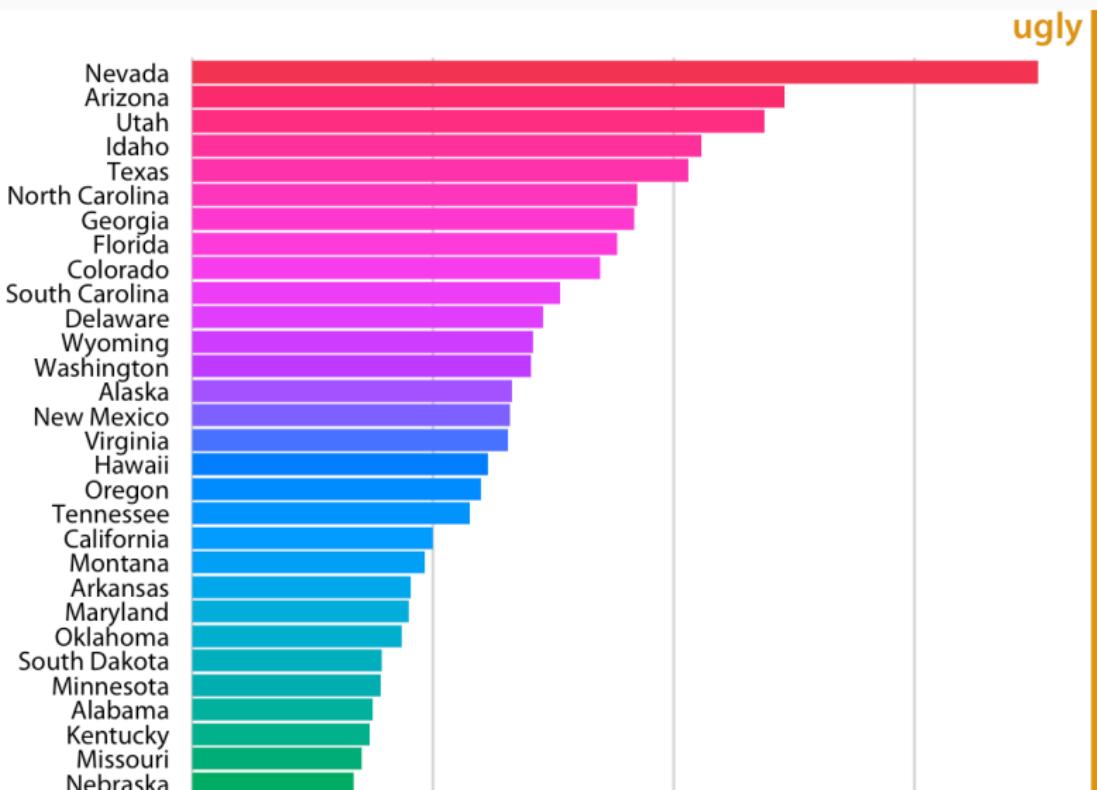


色彩的运用

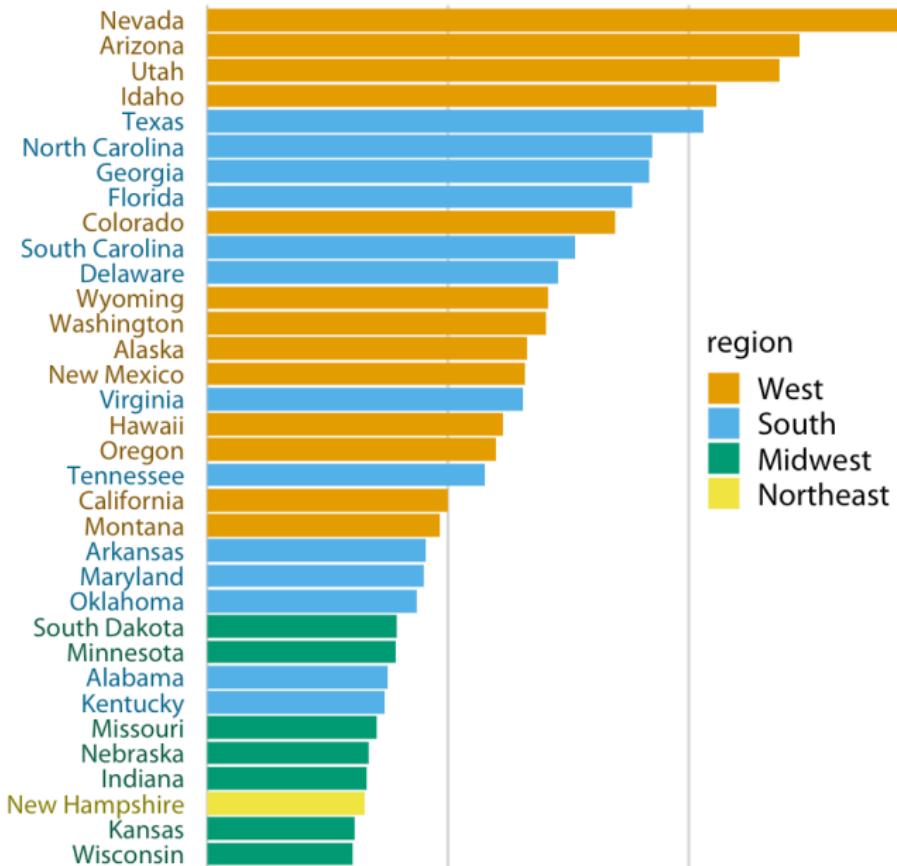


色彩的运用

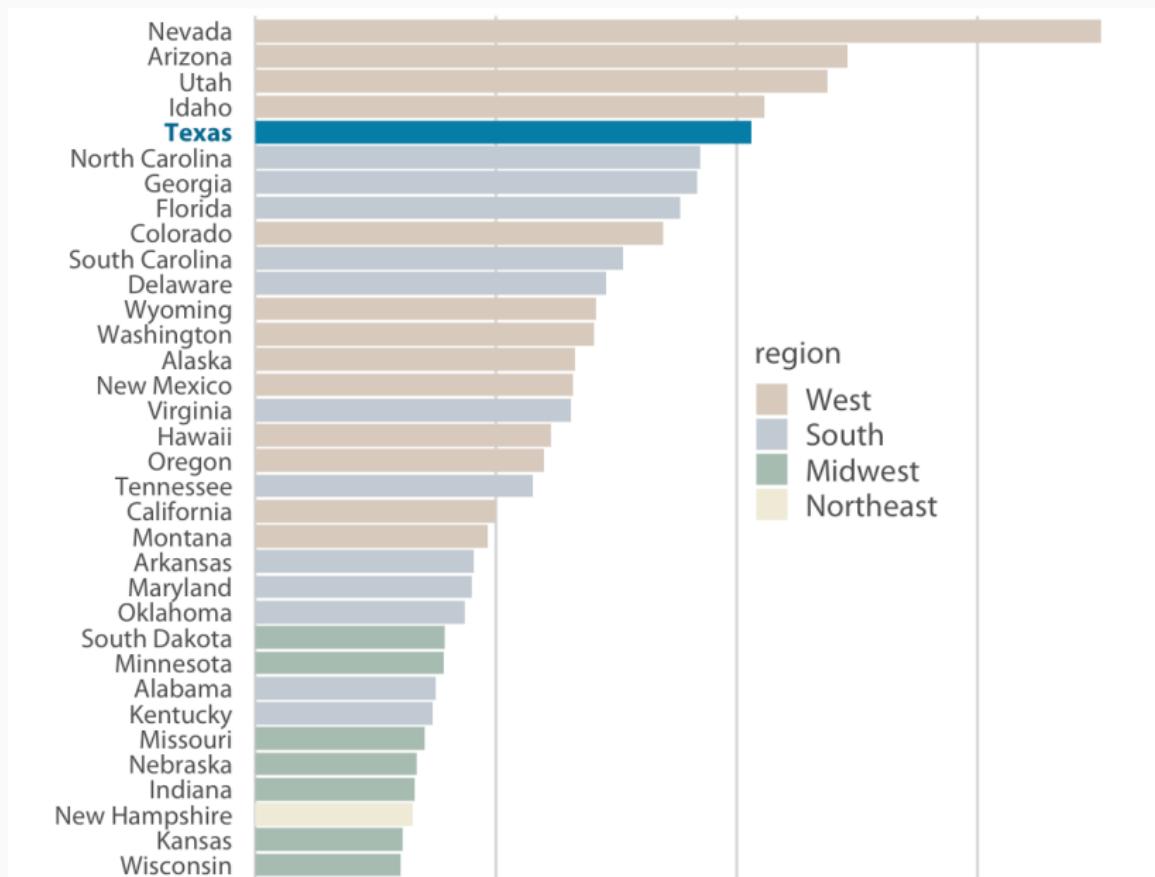
为了颜色而颜色，但一点信息量也没有



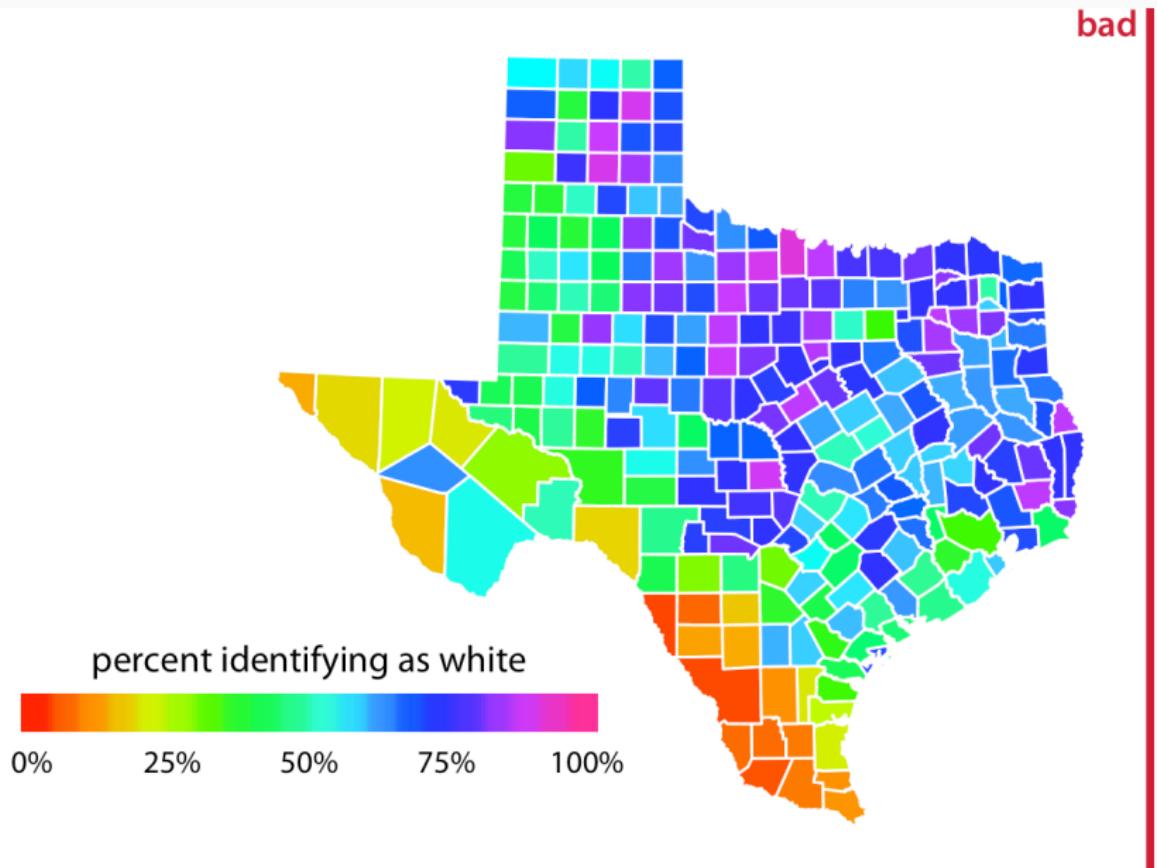
色彩的运用



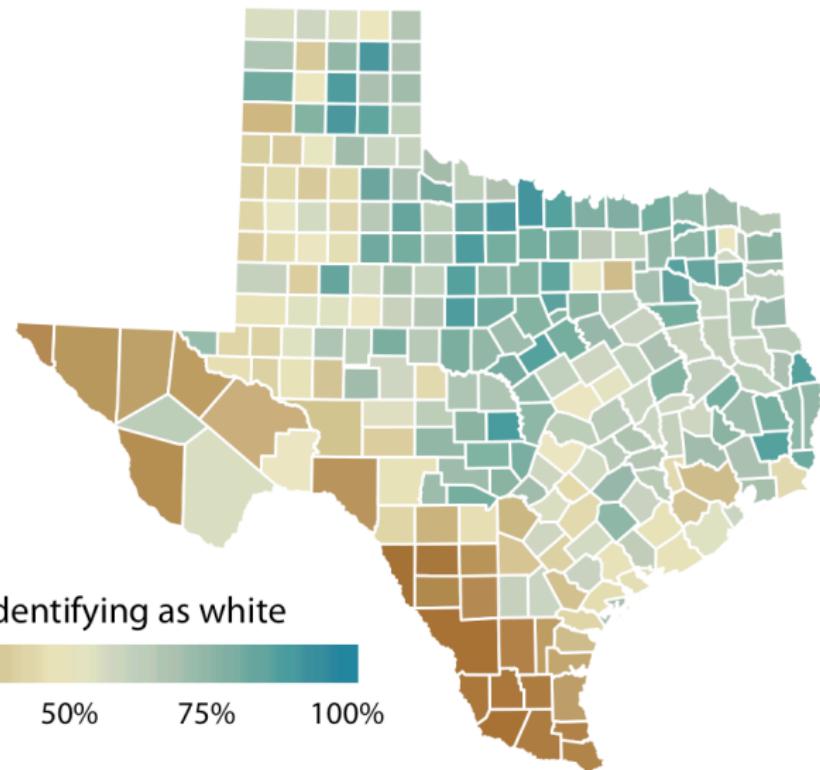
色彩的运用



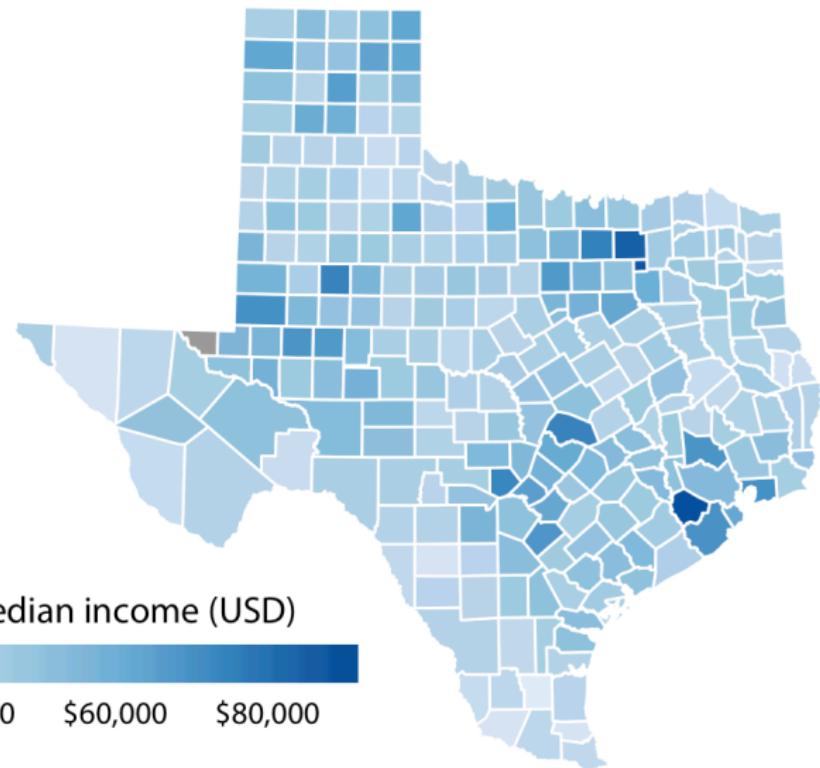
色彩的运用



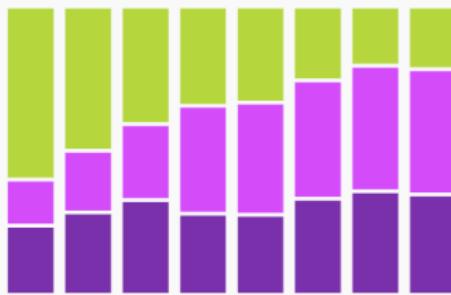
色彩的运用



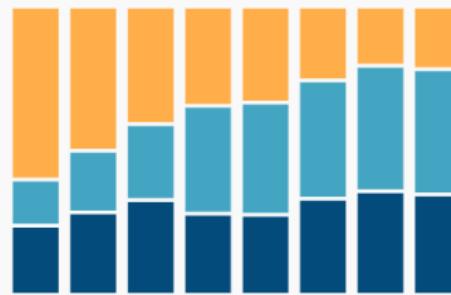
色彩的运用



色彩的运用



NOT IDEAL



BETTER

色彩的运用



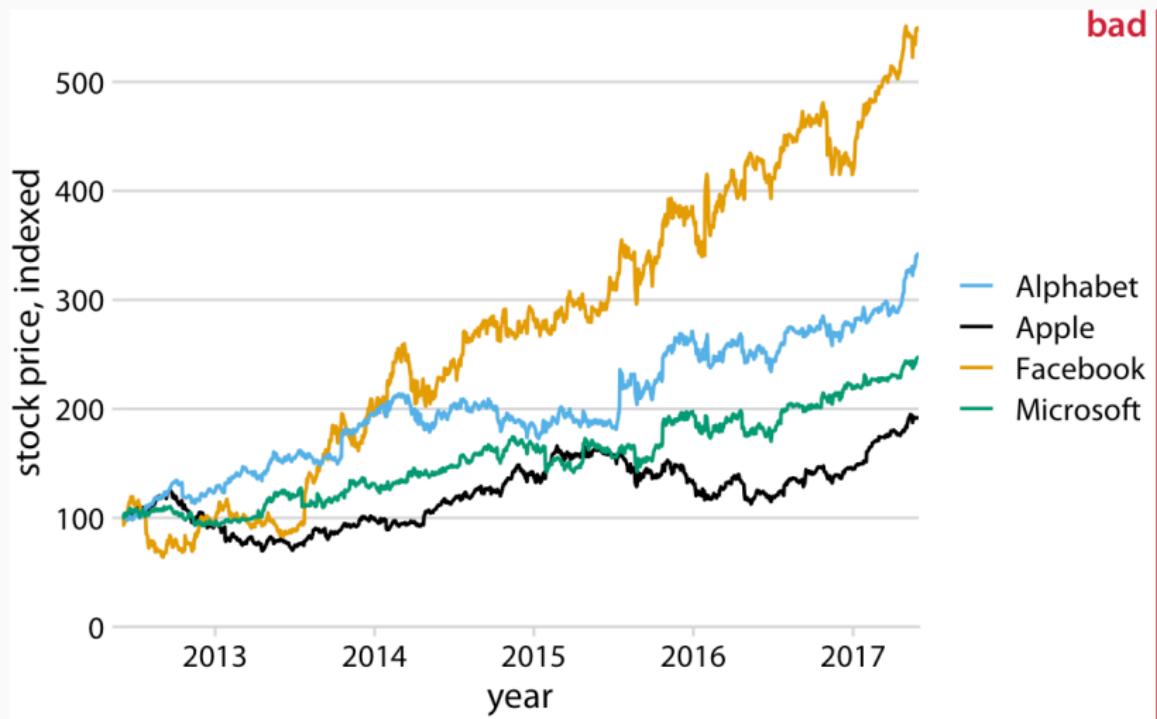
Adobe color

色彩的运用

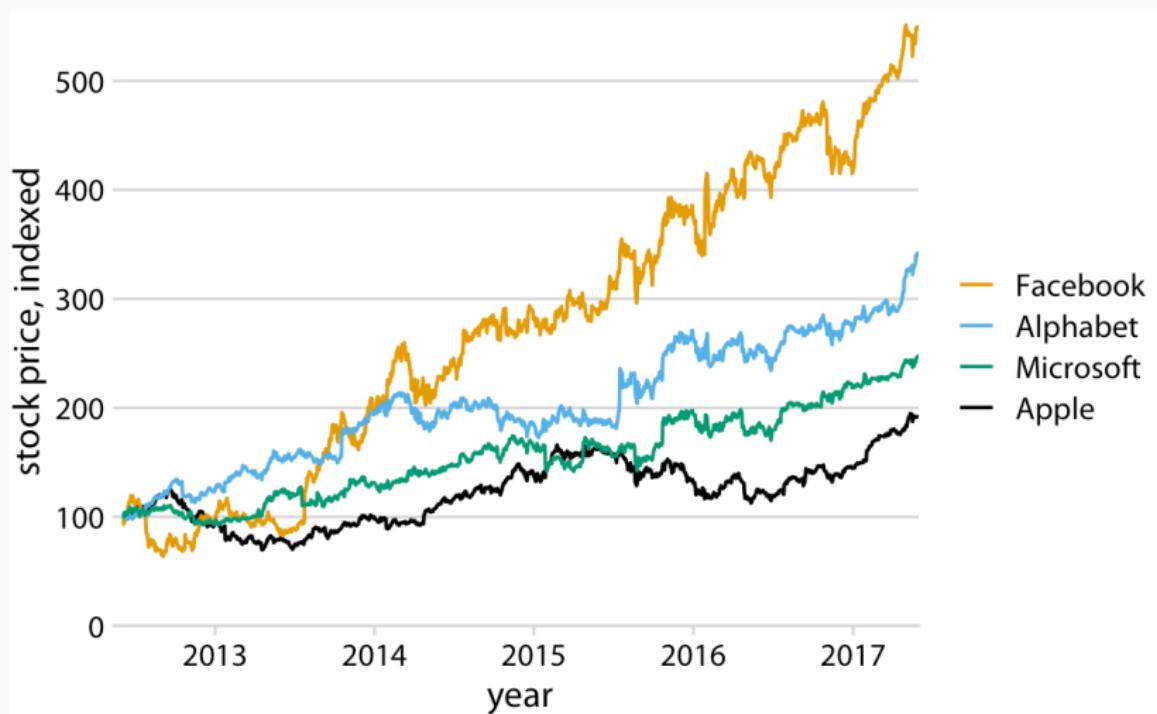


怎么样把配色放到我的图中呢？

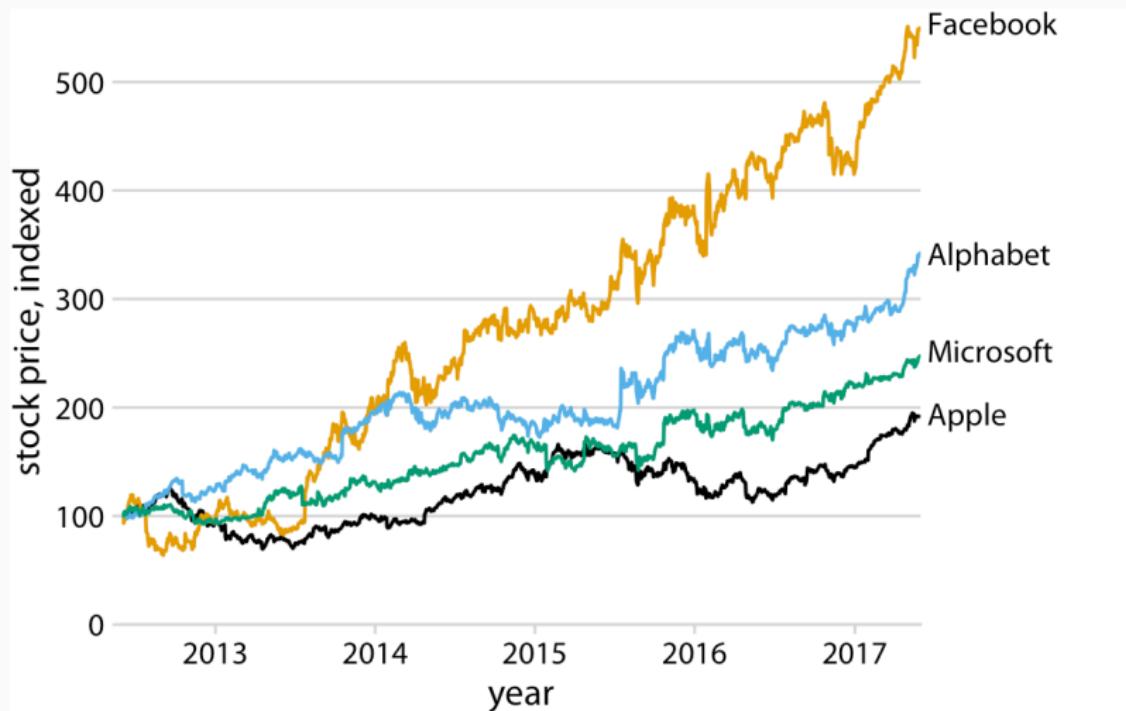
必要的标注



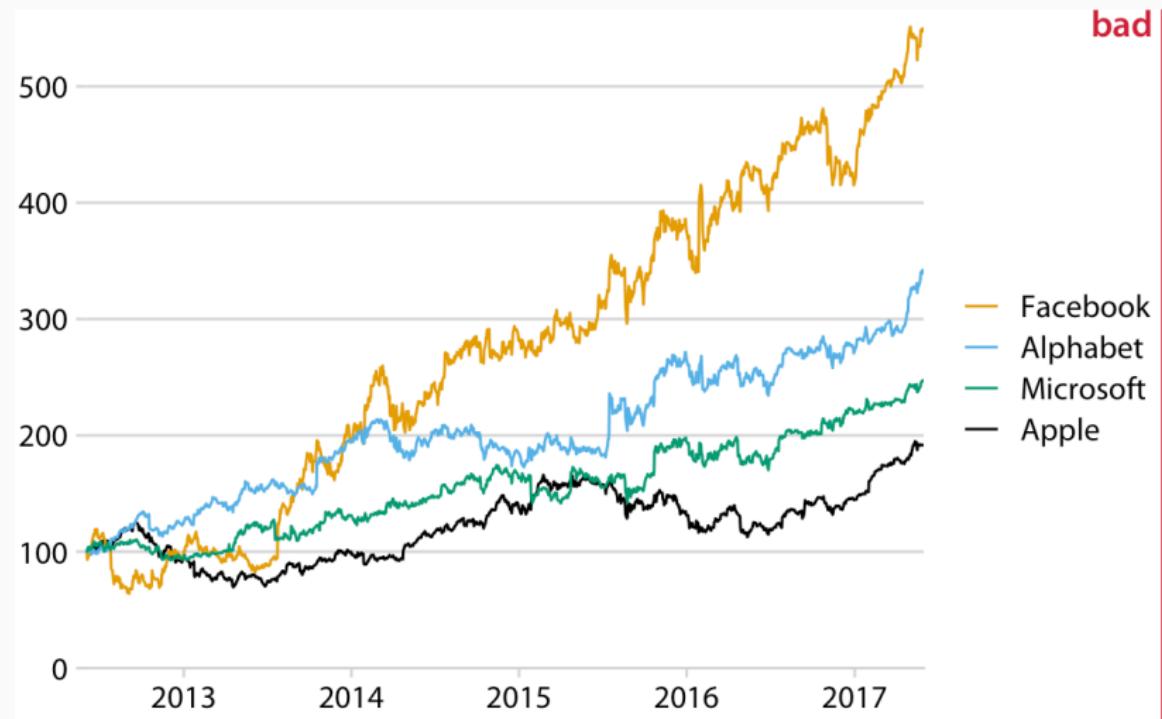
必要的标注



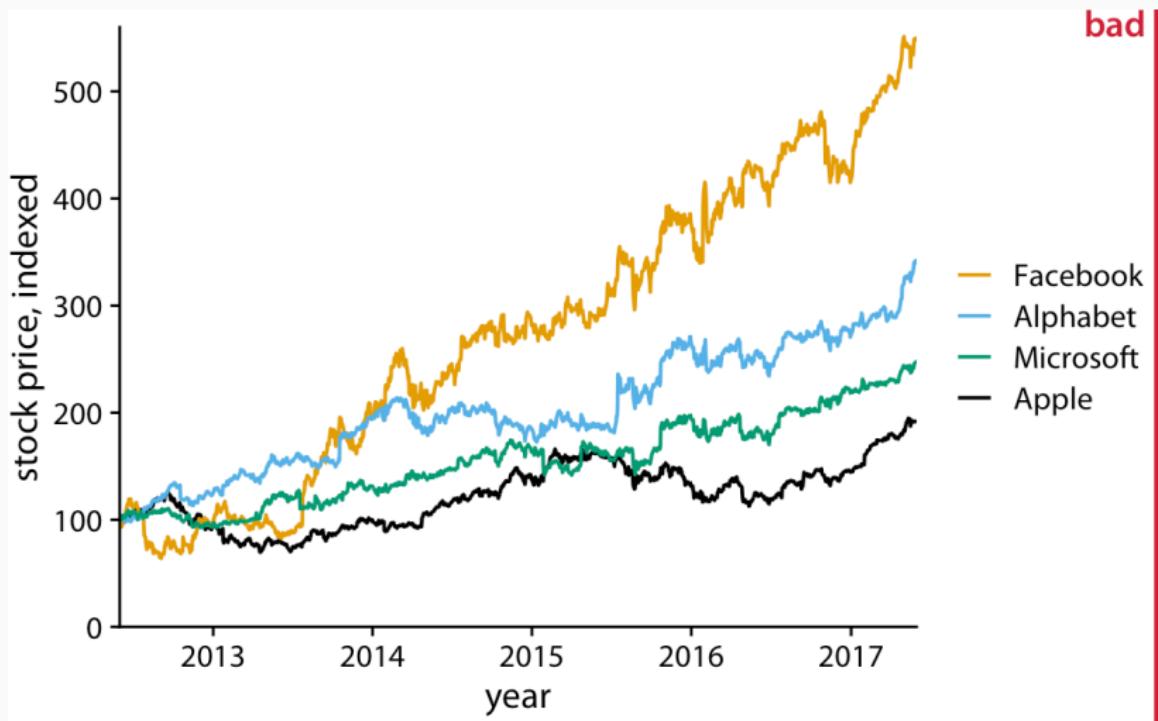
必要的标注



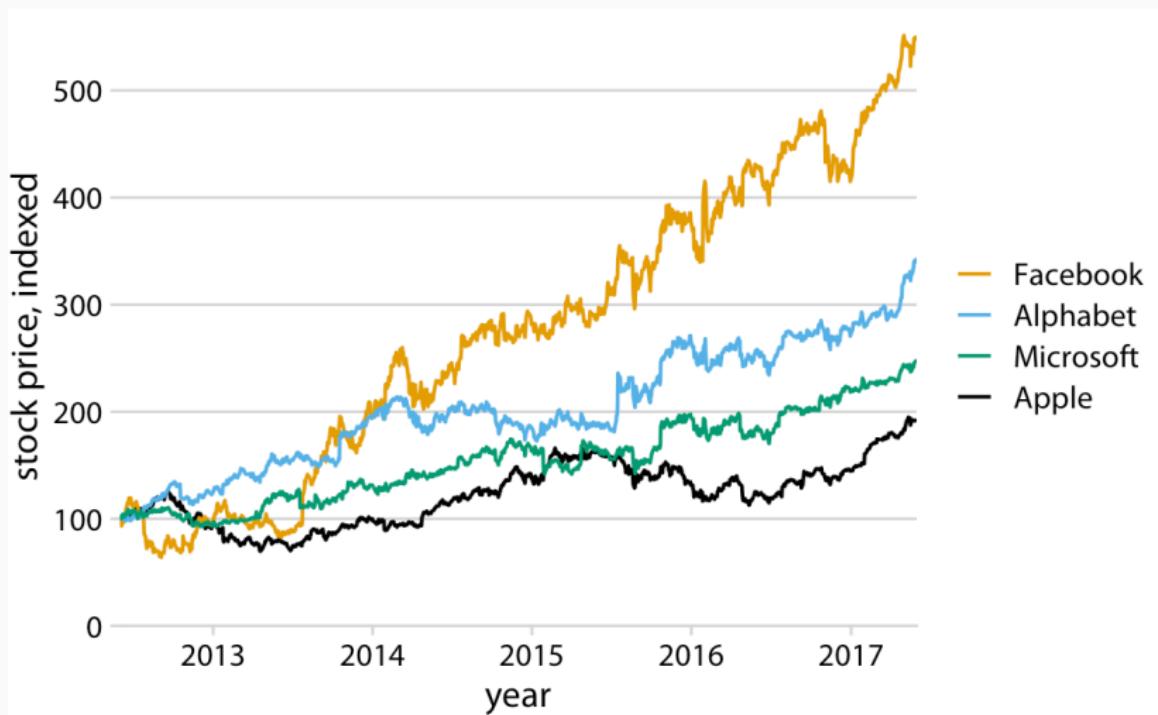
必要的标注



必要的标注



必要的标注

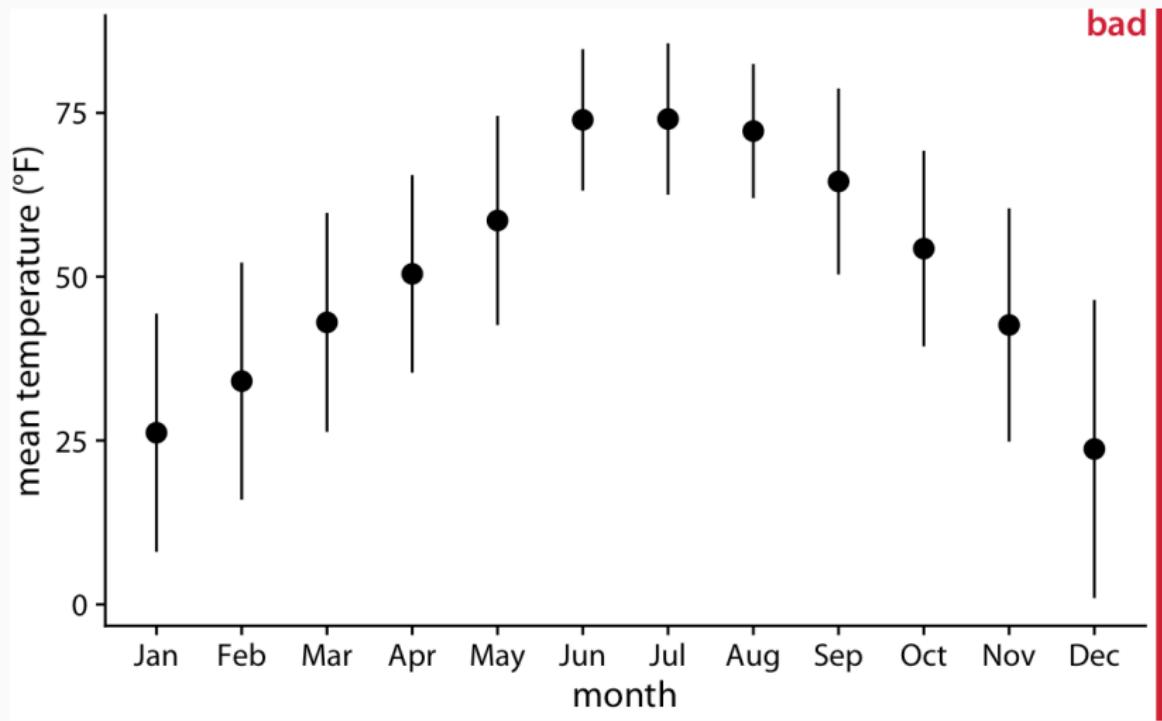


各种分布图的优缺点

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

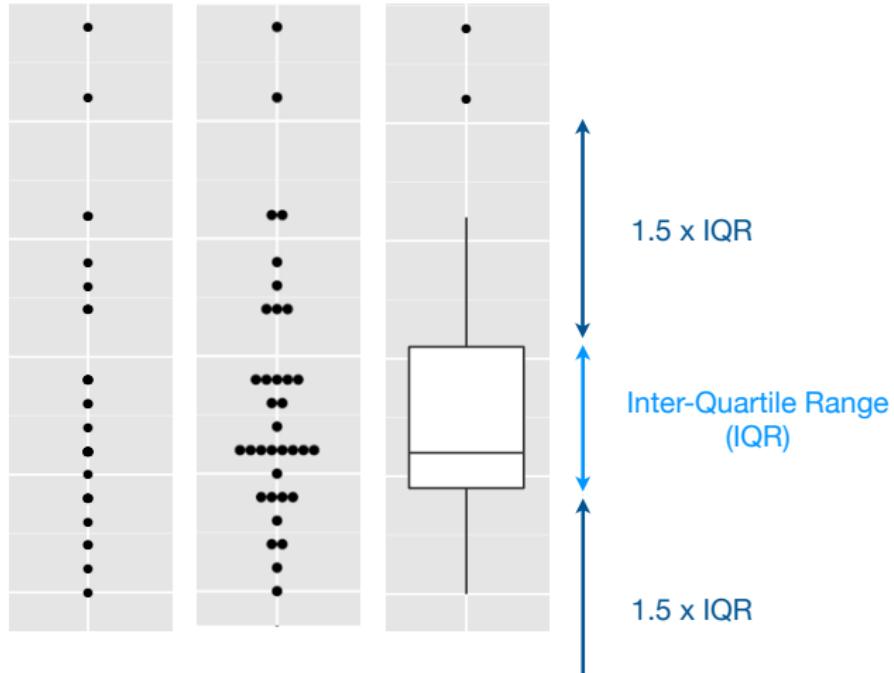
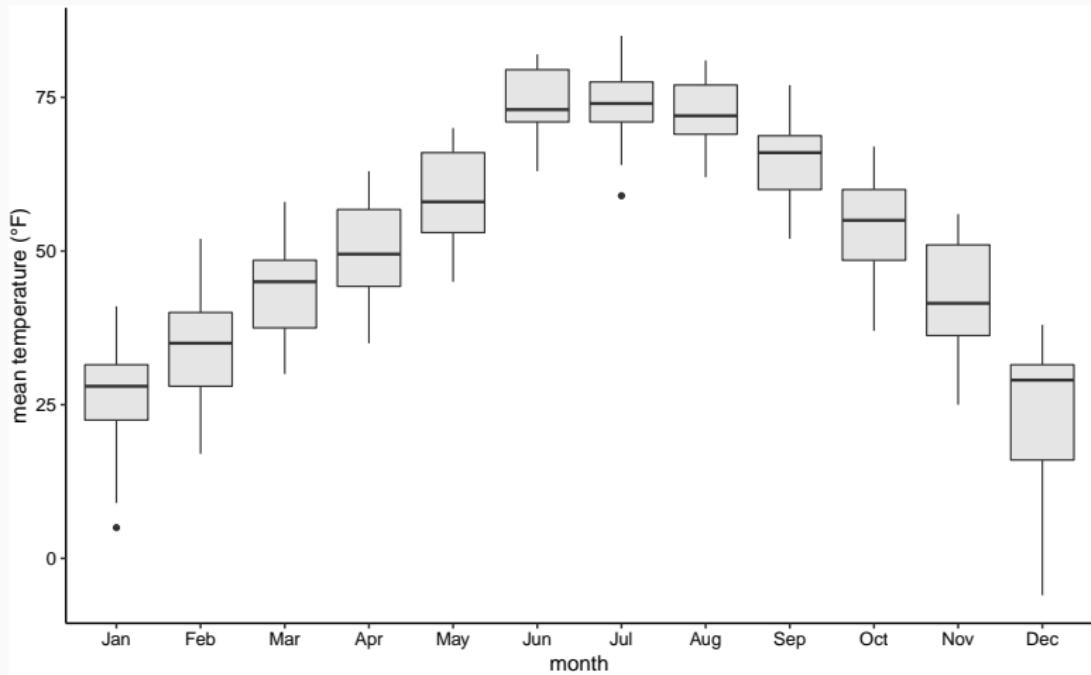
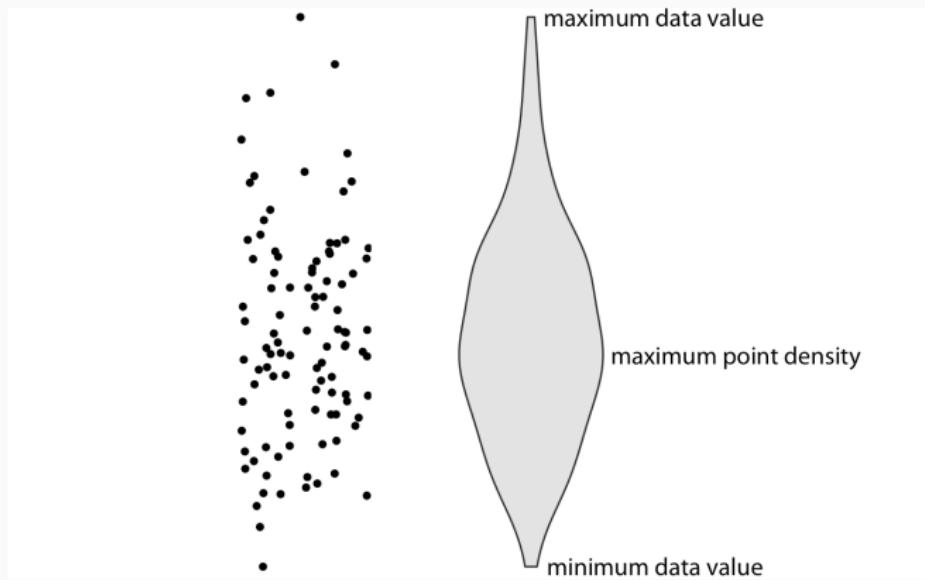


图 1: 箱线图示意图

箱线图



小提琴图



小提琴图的优势

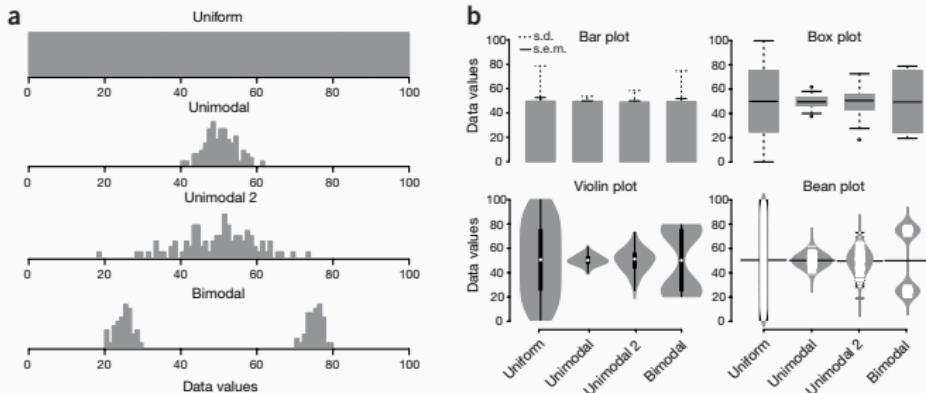
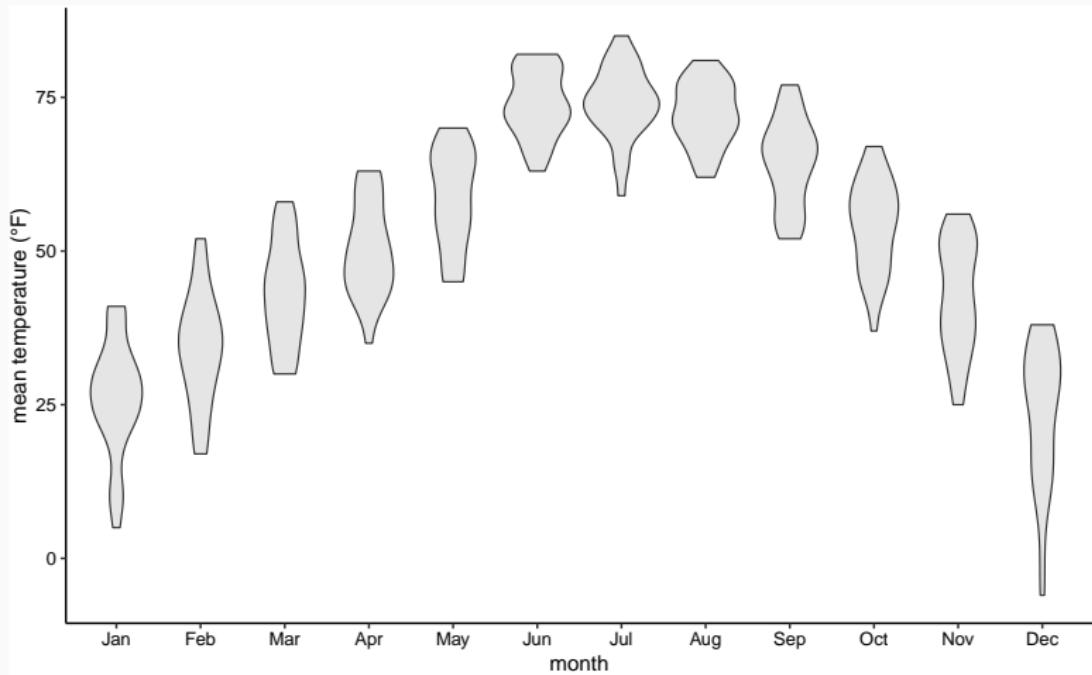


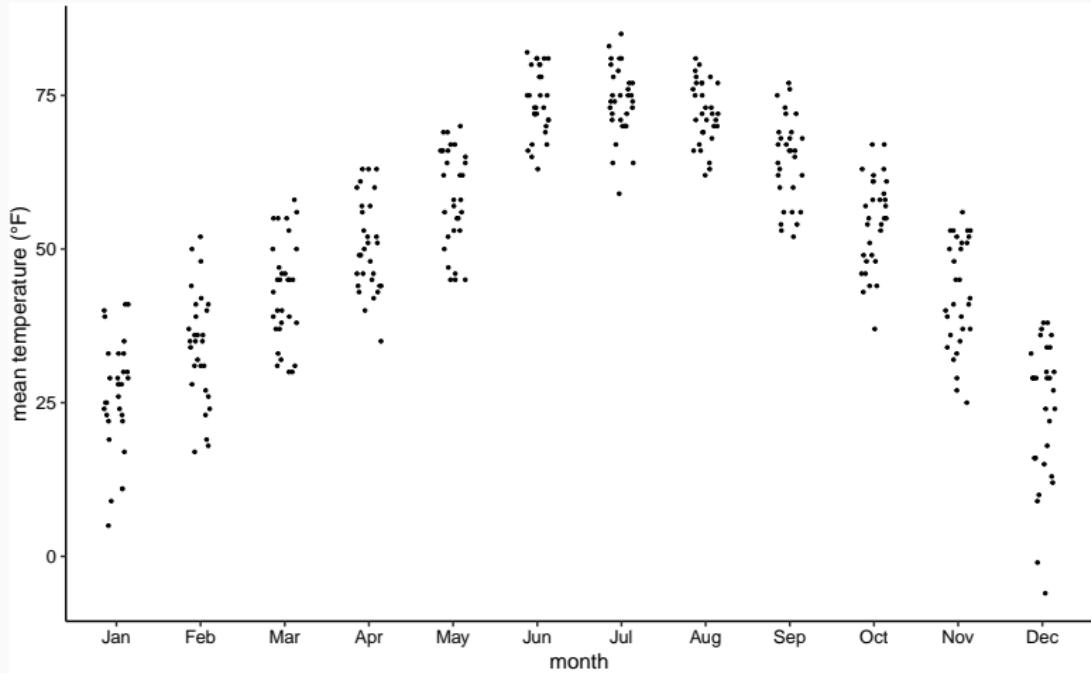
Figure 1 | Data visualization with box plots. (a) Hypothetical sample data sets of 100 data points each that are uniform, unimodal with one of two different variances or bimodal. Simple bar plot representations and statistical parameters may obscure such different data distributions. (b) Comparison of 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

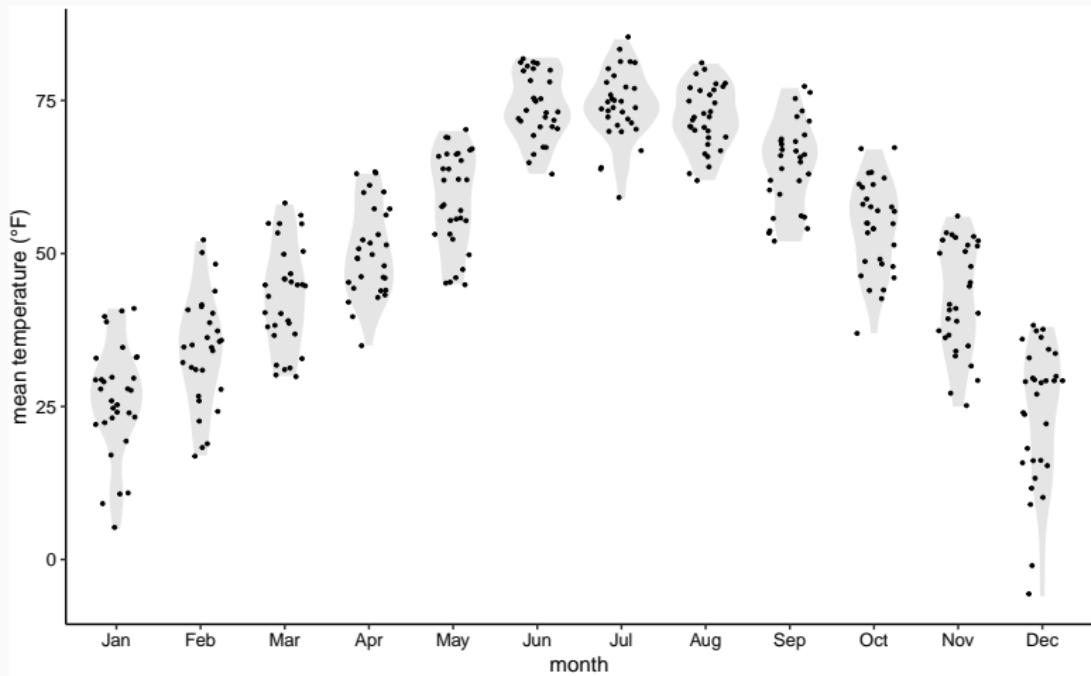
小提琴图



抖散图



Sina 图



山峦图

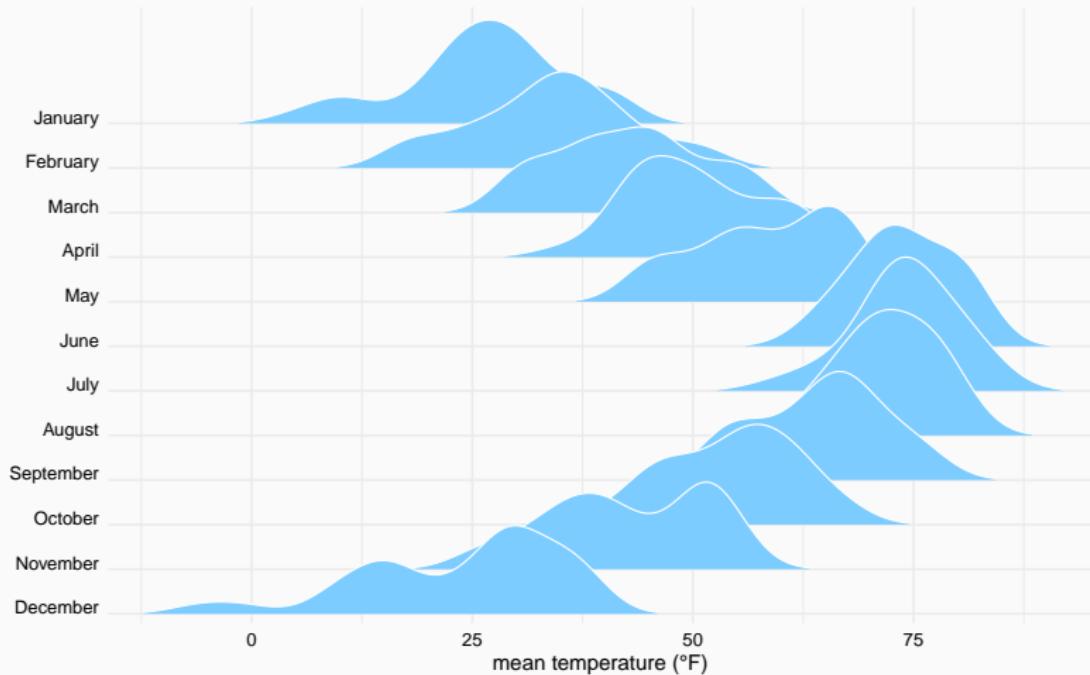


图 3：林肯市 2016 年气温分布山峦图

有颜色山峦图

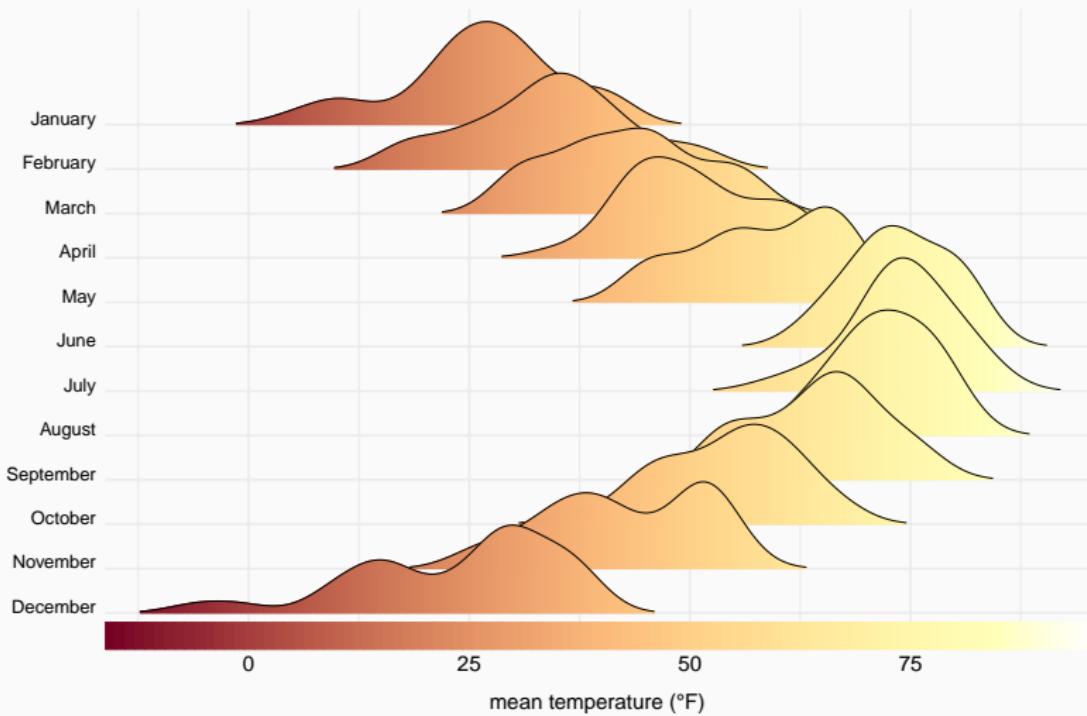
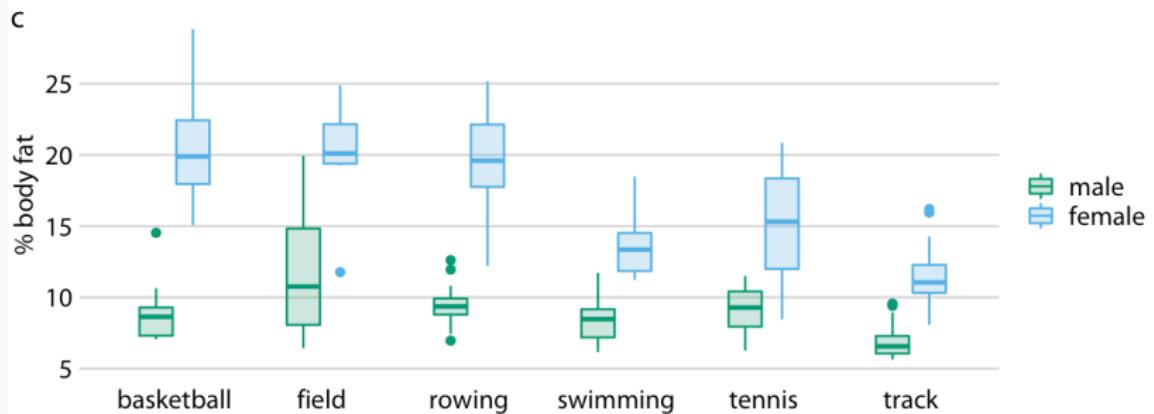
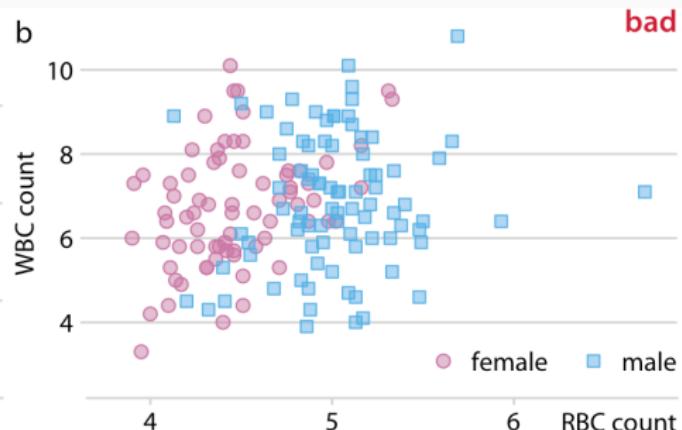
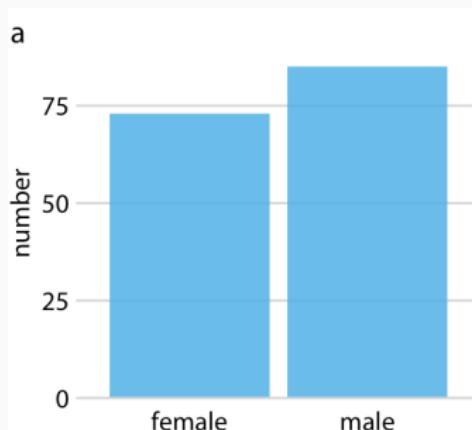
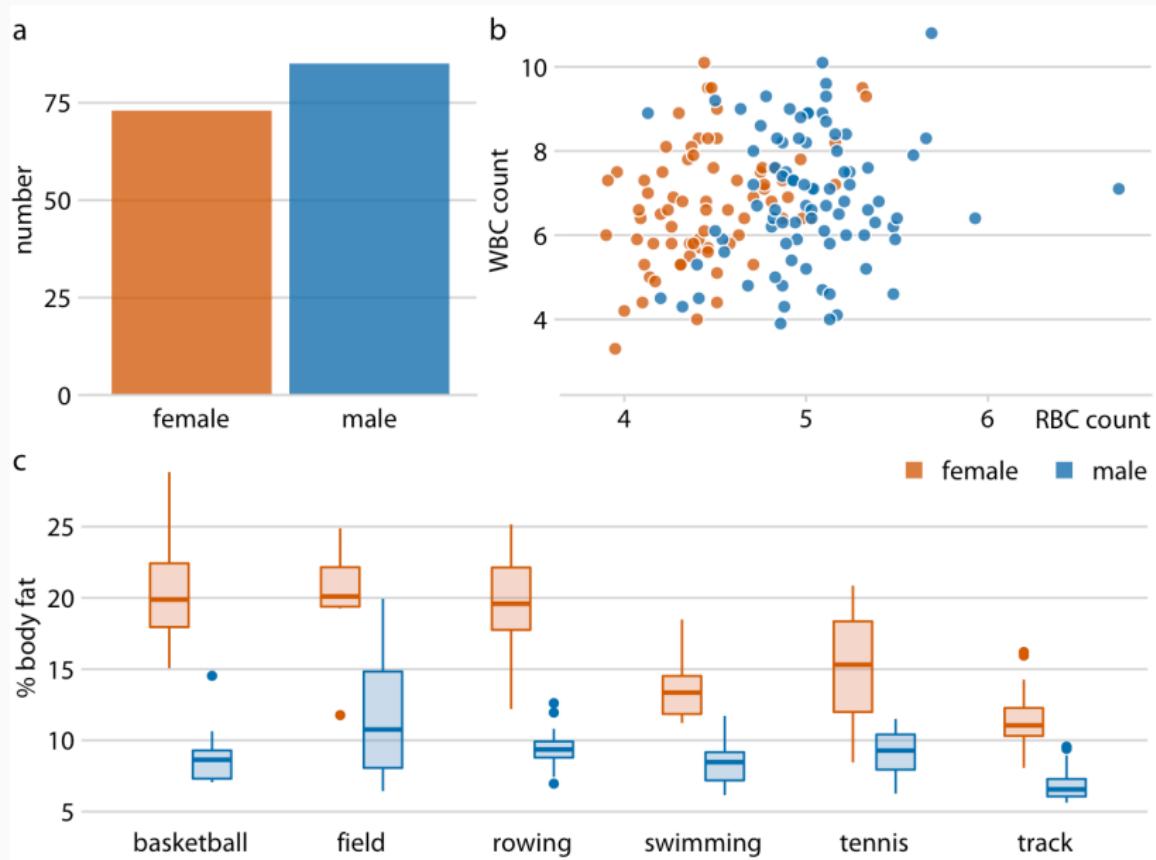


图 4: 林肯市 2016 年气温分布山峦图 (颜色越亮, 温度越高)

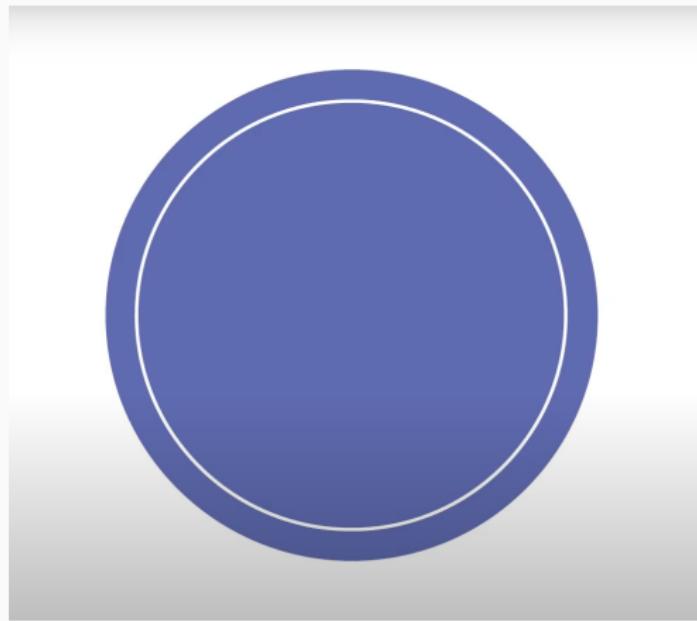
图片组合



图片组合

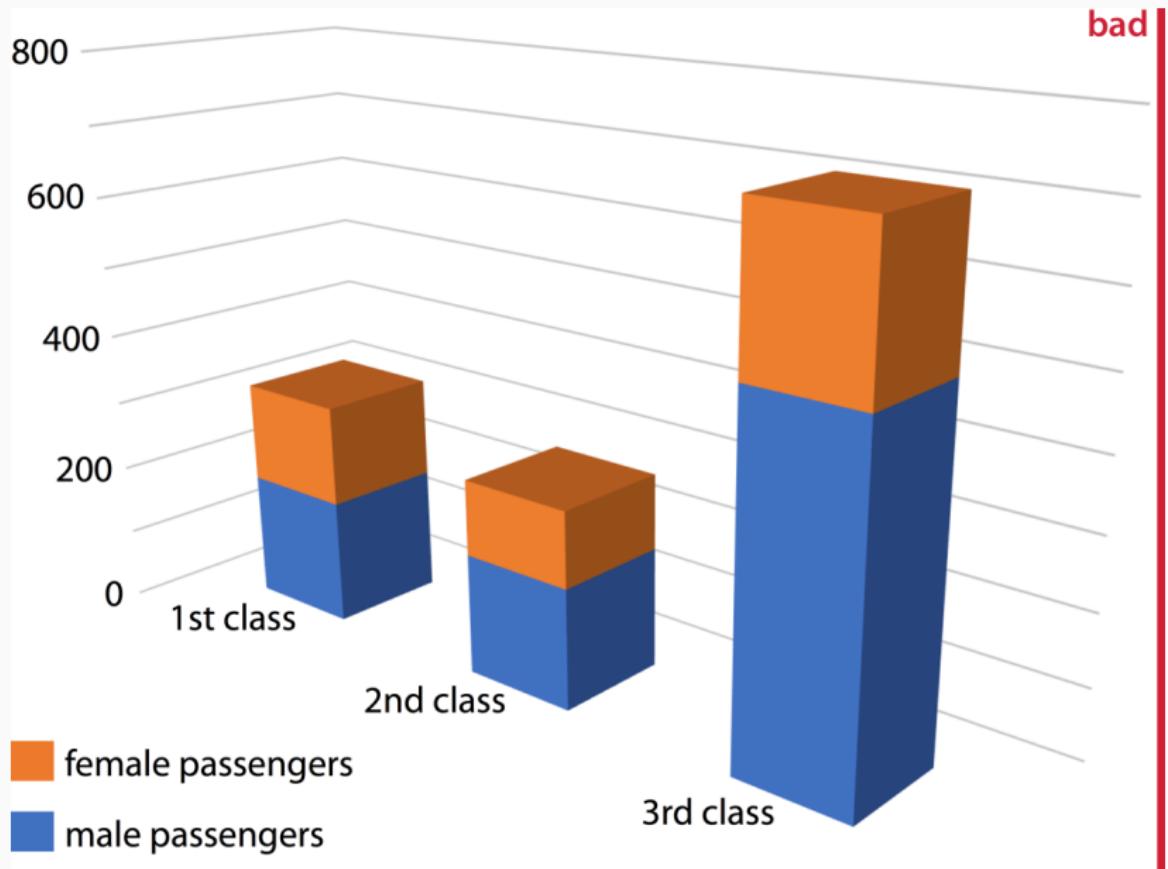


尽可能不用 3D 图

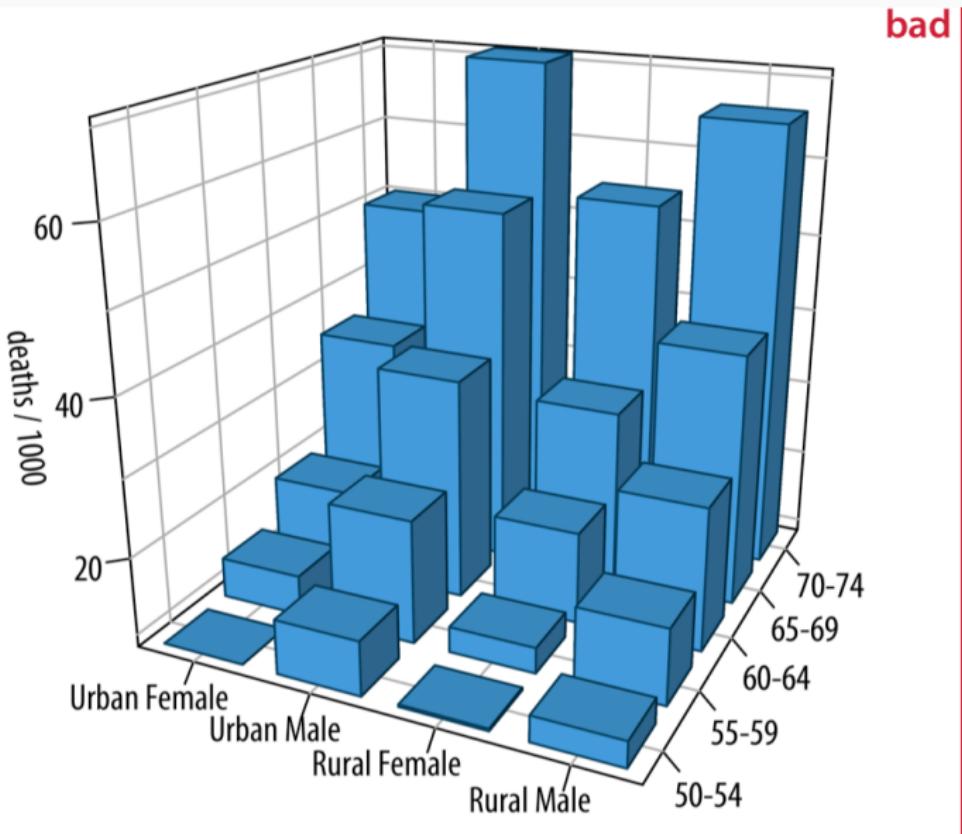


80% - 90% ?

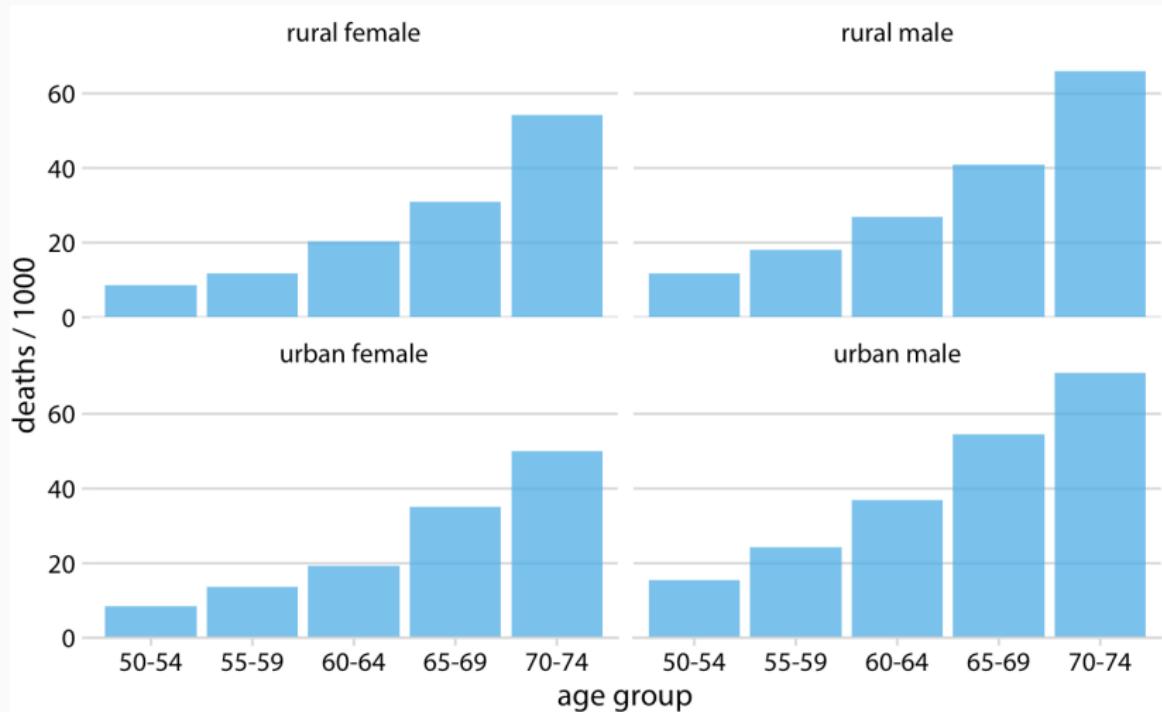
尽可能不用 3D 图



尽可能不用 3D 图



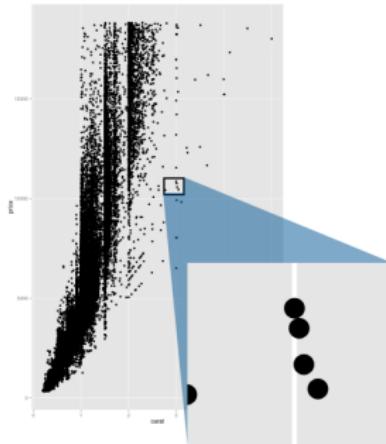
尽可能不用 3D 图



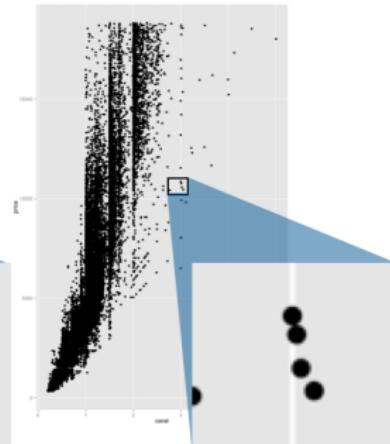
图片格式

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



pdf



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
- 各种分布图的优缺点
- 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



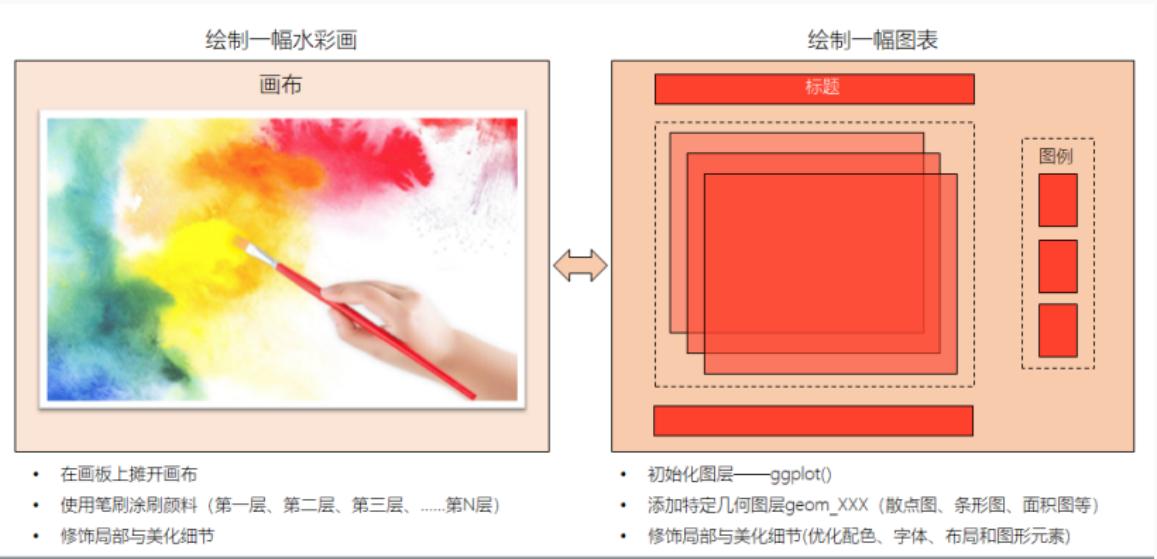
- [Hadley Wickham](#), 一个改变了 R 语言的人

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



ggplot2 语法

```
ggplot(mpg, aes(x = displ, y = hwy)) +  
  geom_point()
```

data set

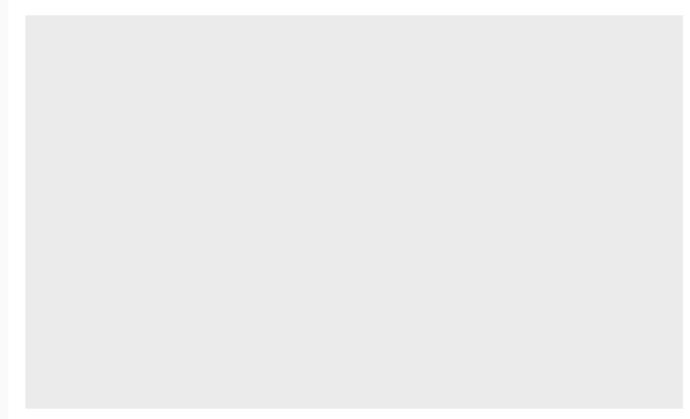
x variable

y variable

"type" of layer

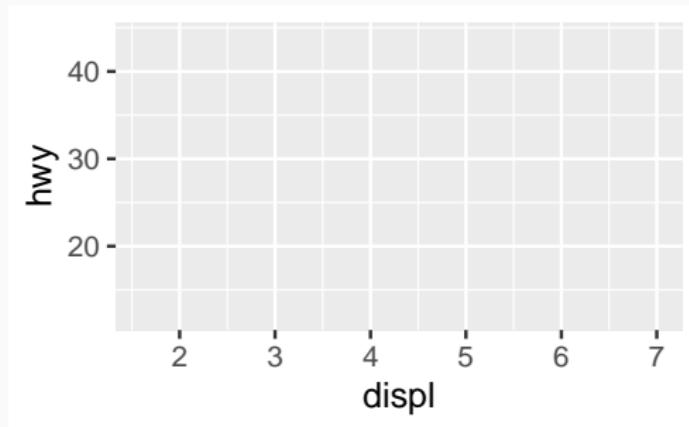
ggplot2 语法很容易理解

ggplot()



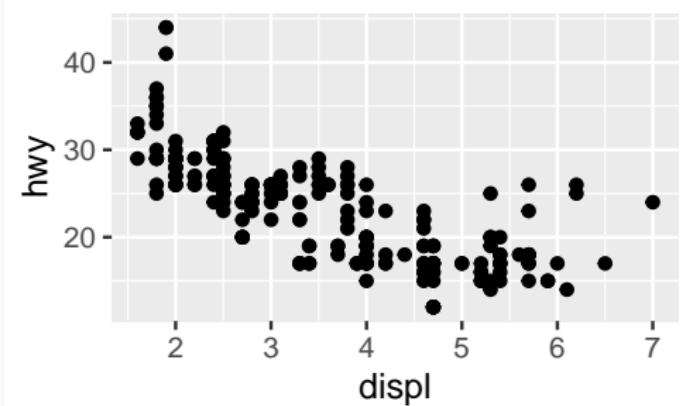
ggplot2 语法很容易理解

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



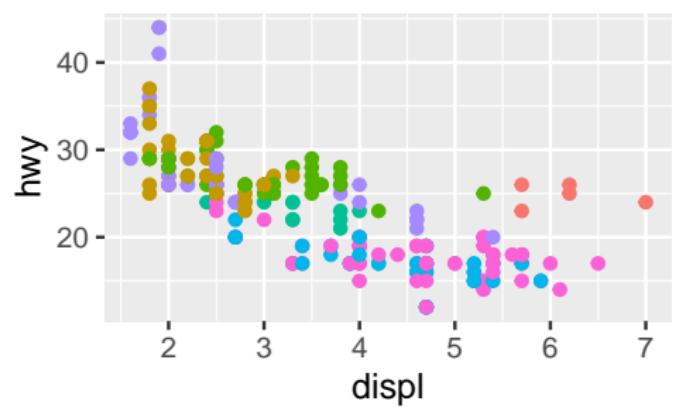
ggplot2 语法很容易理解

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



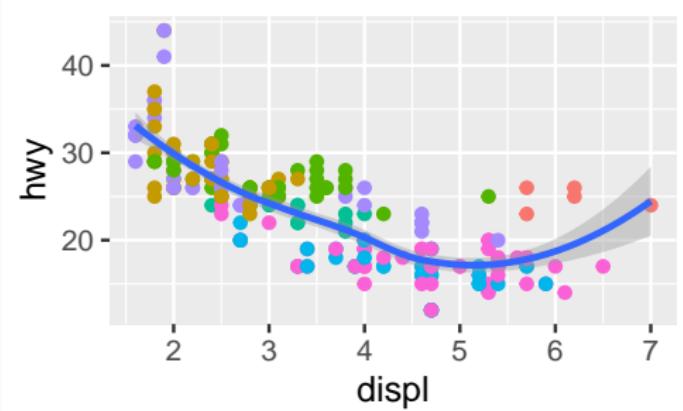
ggplot2 语法很容易理解

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



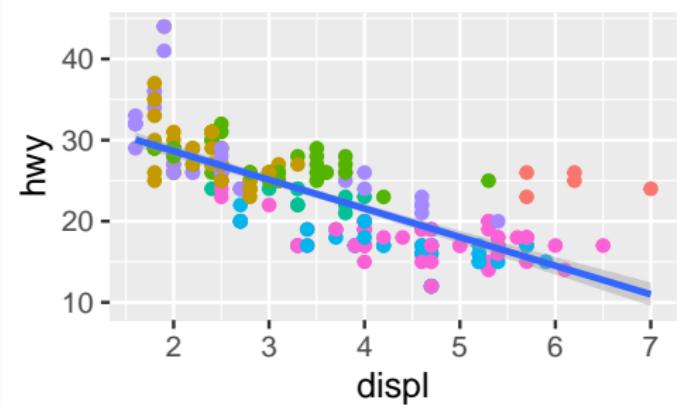
ggplot2 语法很容易理解

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



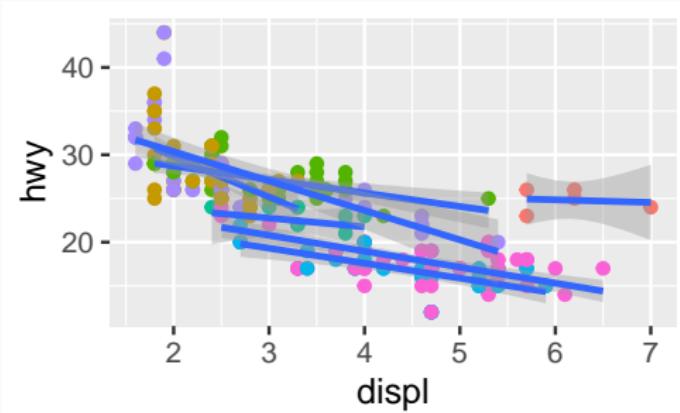
ggplot2 语法很容易理解

```
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

《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。

