北京航空航天大學
经济管理学院
BEIHANG UNIVERSITY
SCHOOL OF ECONOMICS AND MANAGEMENT

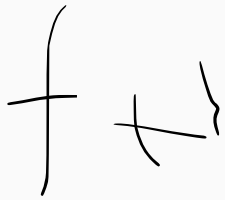
第一讲：描述性统计

康雁飞

2020-04-21

Outline

1. 什么是描述性统计?
2. 汇总统计量
3. 统计图
4. 案例：COVID-19
5. 描述性统计在 R 中的实现
6. 作业



数据分析的基本作为

1. 对系统的描述性分析（现状、结构、因素之间关系等）
2. 对系统的解析性分析（建立模型）
3. 对系统的未来状态进行预测（对未来做出准确的预见）
4. 决策

统计数据的计量尺度

- 定类尺度（Nominal scale）：依据某种尺度进行分类。最粗略、计量程度最低的尺度。
- 定序尺度（Ordinal scale）：分类 + 等级差。
- 定距尺度（Interval scale）：可以测量不同等级之间的差距，分类 + 等级差 + 加减。
- 定比尺度（Ratio scale）：更高一层，分类 + 等级差 + 加减 + 乘除。

统计数据的类型

- 定性数据（Qualitative, Categorical）
  - 名义数据（Nominal data）：分类
  - 顺序数据（Ordinal data）：分类 + 排序
- 定量数据（Quatitative, Numerical）
  - 定距数据（Interval data）：任意两个观测值之间的差值有意义。
  - 定比数据（Ratio data）：任意两个观测值之间的差值和比例值均有意义。

大数据时代的数据类型

多通道、异构数据：文本、图象、音频、视频、位置、轨迹、社交网络

Outline

描述性统计

1. 汇总统计量 - 用少量数字来描述数据（Numerically）
2. 统计图 - 用图将数据可视化（Graphically）

Outline

汇总统计量（summary statistics）

1. 数据的“集中程度”（Central Tendency）
2. 数据的“离散程度”（Spread）

数据的集中程度 - 中位数

将数据从小到大排序后中间位置的观测值。

中位数（Median）

将 n 个数据排序：$x_1 \leq x_2 \leq \cdots \leq x_n$，

1. $\min = x_1$。
2. $\max = x_n$。
3. 中位数为处于中间位置的观测值。

n 为奇数

例：$1, 7, 11, 4, 13$ $(n = 5)$

1. 排序
2. 中位数是几?

n 为偶数

例：$1, 7, 11, 4, 13, 19$（$n = 6$）

1. 排序：1，4，7，11，13，19
2. 中位数是几?
   - 1，4，**7**，**11**，13，19
   - 中位数 $= \frac{7+11}{2} = 9$
   - 中位数是唯一的!

数据的集中程度 - 众数

众数（Mode）

出现频次最多的观测值。

众数的意义

例：服装、鞋帽的生产和销售企业为了掌握消费者的需求和偏好，会关心这些商品的款式、尺码、规格、颜色的众数。

—流行款式、流行色

众数

1 众数不唯一

例：11, 18, 11, 3, 7, 10, 9, 9, 11, 6, 9

众数 = 9，11

2 数据过于分散时，统计众数没有意义

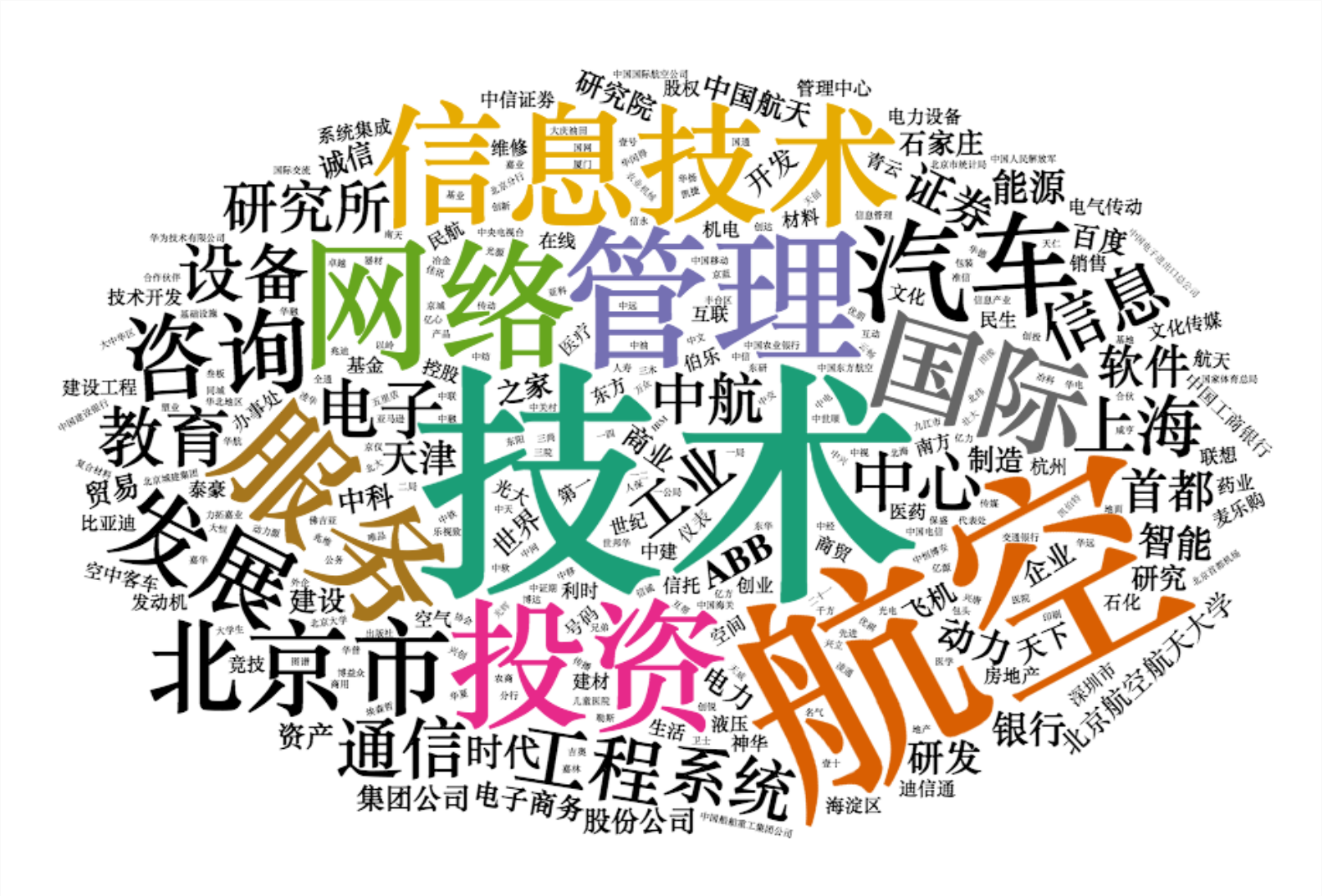
例：某小饭馆在一周内的每日顾客数量：92, 84, 70, 76, 66, 80, 71。

众数 = ?

学历	人数
初中	155
高中	100
大学	185
硕士	55
博士	70
其他	25

词云图

例：MBA 学生所在单位



百度迁徙

来源：http://qianxi.baidu.com/

例：迁入北京来源地

数据的集中程度 - 均值

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

性质 1

观测值与均值的离差和为零。

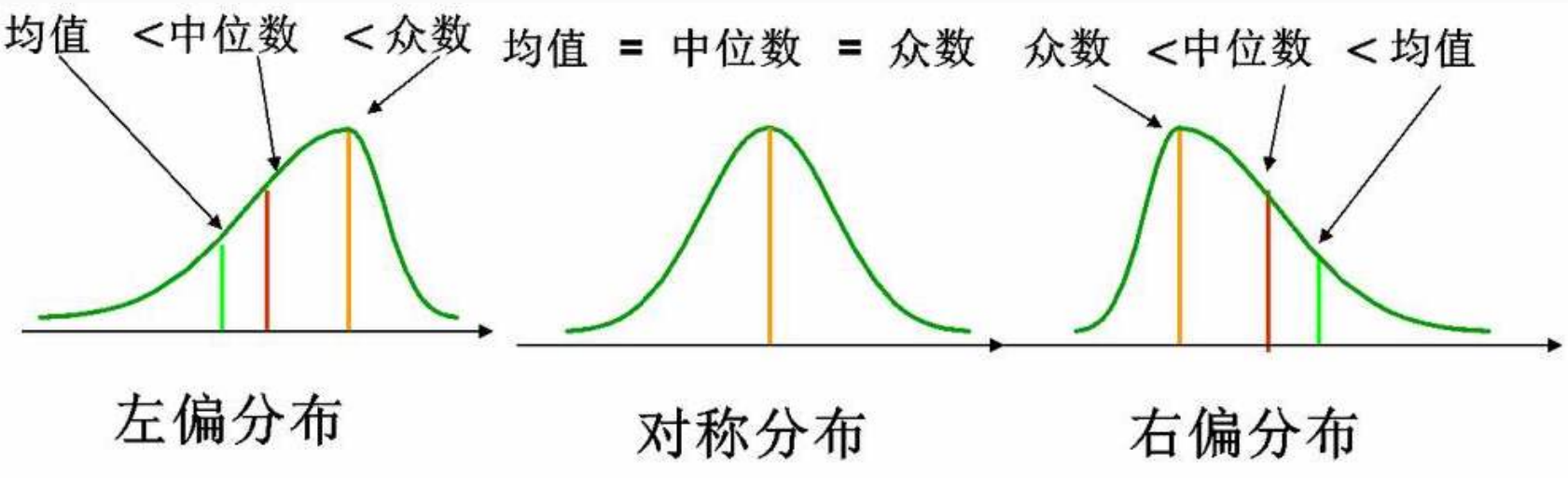
$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

性质 2

观测值与均值的离差的平方和取到最小值。

$$\min_{\alpha \in R}\sum_{i=1}^{n}(x_i - \alpha)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

均值、中位数和众数之间的比较



均值、中位数和众数之间的比较

例：业余歌手大奖赛

11 名裁判给出的得分分别如下。

```
## 9.9 9.3 9.3 9.3 9.2 8.9 8.8 8.8 8.7 8.5 8.4
```

```
## mean =  9.009091
```

```
## median =  8.9
```

```
## mode =  9.3
```

均值、中位数和众数之间的比较

例：业余歌手大奖赛-复赛

```
## 9.9 9.3 9.3 9.2 9.2 8.8 8.8 8.8 8.7 8.5 8.4
```

```
## mean =  8.990909
```

```
## median =  8.8
```

```
## mode =  8.8
```

均值、中位数和众数之间的比较

例：业余歌手大奖赛-异常值

```
## 9.9 9.3 9.3 9.2 9.2 8.8 8.8 8.8 8.7 8.5 0.4
```

```
## mean =  8.263636
```

```
## median =  8.8
```

```
## mode =  8.8
```

比较：均值、中位数和众数

均值	中位数	众数
数据全部信息	中间位置数据	峰值
具有唯一性	具有唯一性	不具有唯一性
比较稳定	比较稳定	最不稳定
受极端值影响	对极端值不敏感	不受极端值影响

应用建议

1. 对于名义变量，描述集中趋势的办法是：众数。
2. 对于顺序变量，描述集中趋势的常用的办法是：中位数。
3. 对于定量变量，一般使用平均值。
4. 但均值受极端值影响，当数据非对称分布时，可用中位数。

汇总统计量（summary statistics）

1. 数据的“集中程度”（Central Tendency）
2. 数据的“离散程度”（Spread）

数据的离散程度

例：某车间有 A、B 两人，加工某零件质量（直径，mm）

- A：0.7，0.7，0.8，0.8，0.8，0.8，1.0，1.1
- B：0.4，0.5，0.7，0.8，0.8，0.9，1.2，1.4

数据的离散程度-极差

- A：Range = 1.1 - 0.7 = 0.4
- B：Range = 1.4 - 0.4 = 1.0

极差（Range）

- 最大值与最小值之间的差距
- 极差容易受极端值的影响

数据的离散程度-四分位极差

四分位极差（Interquantile Range）

将 n 个数据排序：$x_1 \leq x_2 \leq \cdots \leq x_n$，

- Q_1 为低四分位数，lower quantile，25% percentile.
- Q_3 为高四分位数，upper quantile，75% percentile.
- 四分位极差 = Q_3 - Q_1.

2 月 9 日，medRxiv 上由钟南山院士团队发表的论文《中国 2019 年新型冠状病毒感染的临床特征》，分析了全国 30 个省市 552 家医院 1099 例患者临床特征。

“潜伏期最短为 0 天，最长可达到 24 天；中位数据为 3 天；最长的 24 天（1 人）；四分位距，分别是 2 天与 7 天。”

— 钟南山院士团队《中国 2019 年新型冠状病毒感染的临床特征》

数据的离散程度-方差

如何全面测度观测点集合的差异程度?

方差（Variance）

思路：以均值为中心，测量所有观测值与均值的平均偏离程度。

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2.$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

它们有什么区别?

标准差（Standard Deviation）

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}.$$

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- 优点:与原变量同尺度（scale)。
- 如果两组数据相差较大，是否可以直接比较它们的方差或标准差?

离散系数（Coefficient of Variation，CV）

$$CV = \frac{S}{\bar{x}}.$$

均值相同或相近时，才能用方差或标准差比较数据的离散程度。

1. 4，5，6，7，8
2. 40，50，60，70，80

Outline

定性变量

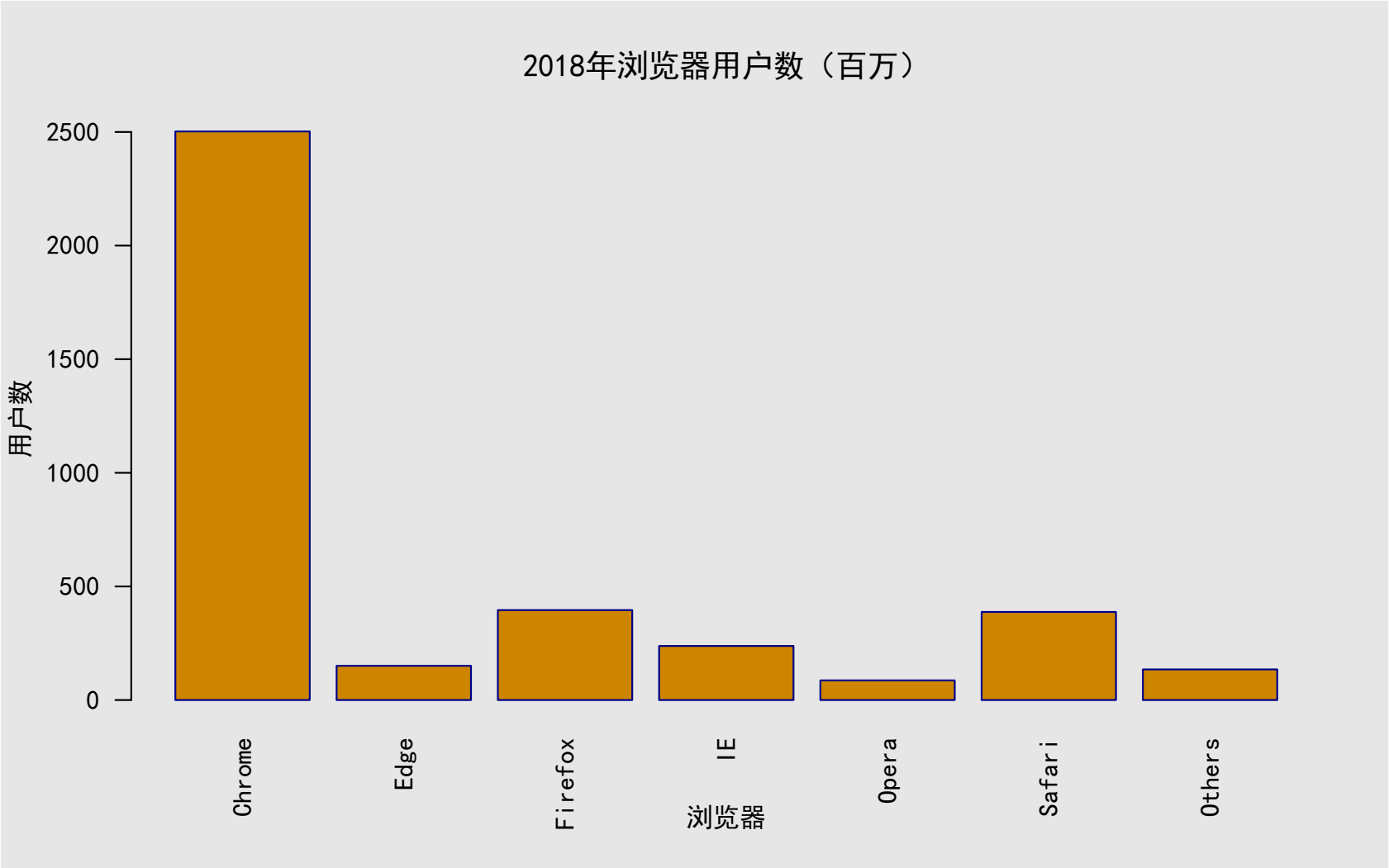
- 条形图
- 饼图

条形图（Bar chart）

下边列出 2018 年每个浏览器的用户数（单位：百万；来源：statista.com）

```
##     浏览器   用户数
## 1  Chrome 2502.40
## 2    Edge  150.78
## 3 Firefox  395.83
## 4      IE  238.05
## 5   Opera   86.49
## 6  Safari  387.65
## 7  Others  134.80
```

条形图


2018年浏览器用户数（百万）
用户数
2500
2000
1500
1000
500
0
Chrome
Edge
Firefox
IE
Opera
Safari
Others
浏览器

豆瓣电影评分

豆瓣电影 搜索电影、电视剧、综艺、影人

影讯&购票 选电影 电视剧 排行榜 分类 影评 2019年度榜单 2019书影音报告

豆瓣 2019 年度电影榜单

姐妹老板 Like a Boss (2020)

导演: 米盖尔·阿尔特塔
编剧: 萨姆·皮特曼 / 亚当·科尔·凯利
主演: 萨尔玛·海耶克 / 罗丝·伯恩 / 蒂凡尼·哈迪斯 / 比利·波特 / 詹妮佛·库里奇 / 更多...
类型: 喜剧
制片国家/地区: 美国
语言: 英语
上映日期: 2020-01-10(美国)
片长: 83分钟
又名: 有限合伙人 / Limited Partners / 波士鬥腦細(港)
IMDb链接: tt7545266

豆瓣评分
5.0 318人评价

5星 2.5%
4星 6.0%
3星 43.1%
2星 36.7%
1星 11.7%

想看 看过 评价: ☆☆☆☆☆

写短评 写影评 分享到 推荐

姐妹老板的剧情简介 · · · · · ·

好朋友米娅和梅尔(哈迪斯、萝丝饰)经营着自己白手起家的化妆品公司，过着最好的生活。不幸的是，他们在财务上陷入困境，而化妆品行业臭名昭著的巨头克莱尔(海耶克饰)提出的巨额收购要约太过诱人，这让俩人的终身友谊岌岌可危。

豆瓣成员常用的标签 · ·

喜剧 美国 女性 2
剧情 2019 派拉蒙

以下豆列推荐 · · · · · · (

调侃是一种境界 (duckducker
2020年所有值得关注的电影 (
自留【2015~】 (13yi)
众里等片千百度~~~2019 (Ha
Indie Movie Releases 已有资

谁在看这部电影 · · · · ·

發
今天上午 想看

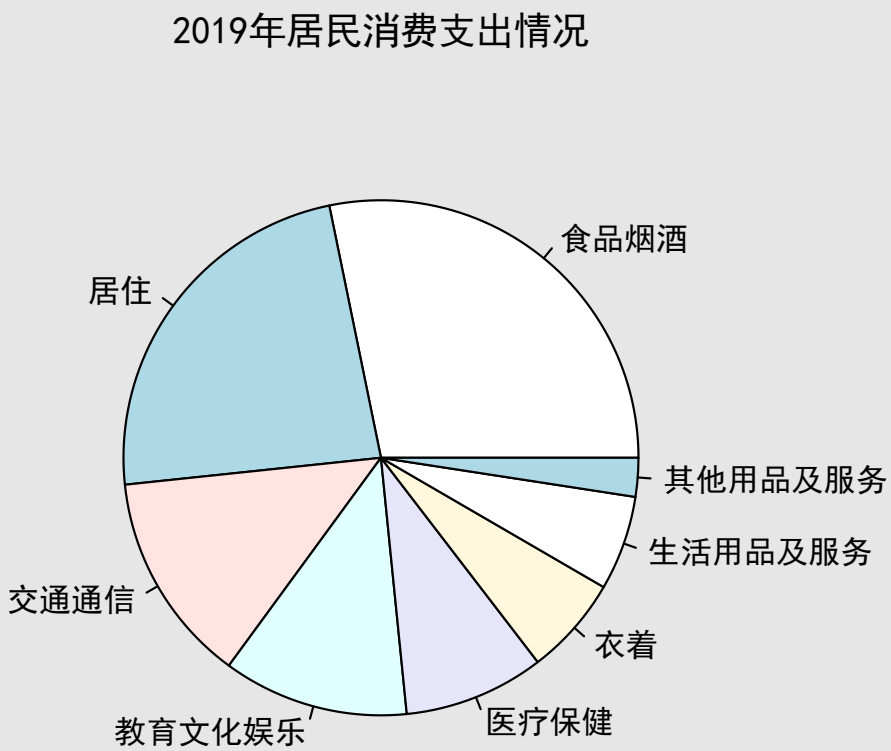
饼图（Pie chart）

例：2019 年居民消费支出情况（来源：国家统计局网站）

```
##    消费                 类别
## 1 6084           食品烟酒
## 2 5055               居住
## 3 2862           交通通信
## 4 2513   教育文化娱乐
## 5 1902           医疗保健
## 6 1338               衣着
## 7 1281 生活用品及服务
## 8  524 其他用品及服务
```

饼图

适于表示数据的结构性特征。



定量变量

- 直方图
- 箱线图
- 线图
- 雷达图
- 散点图
- 热力图
- 气泡图
- 地图

直方图（Histogram）

例：全班 50 名同学的统计学考试成绩

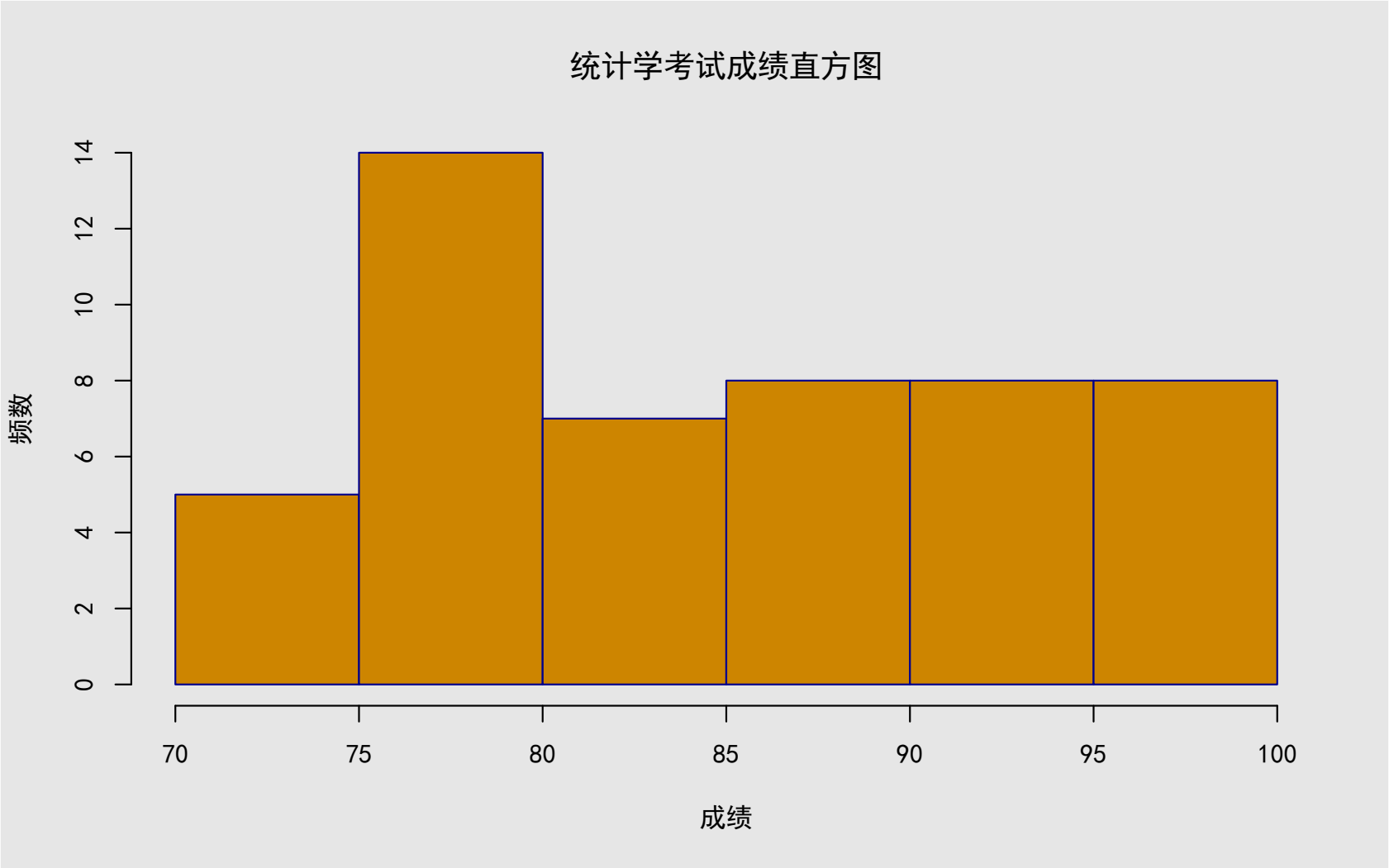
```
##  [1] 100  84  88  83  72  79  87  91  80  74
## [20]  94  97  94  78  98  72  77  95  76  79
## [39]  81  84  79  82  76  78  78  79  92  96
```

Q1：你从这些数据中得出了什么？

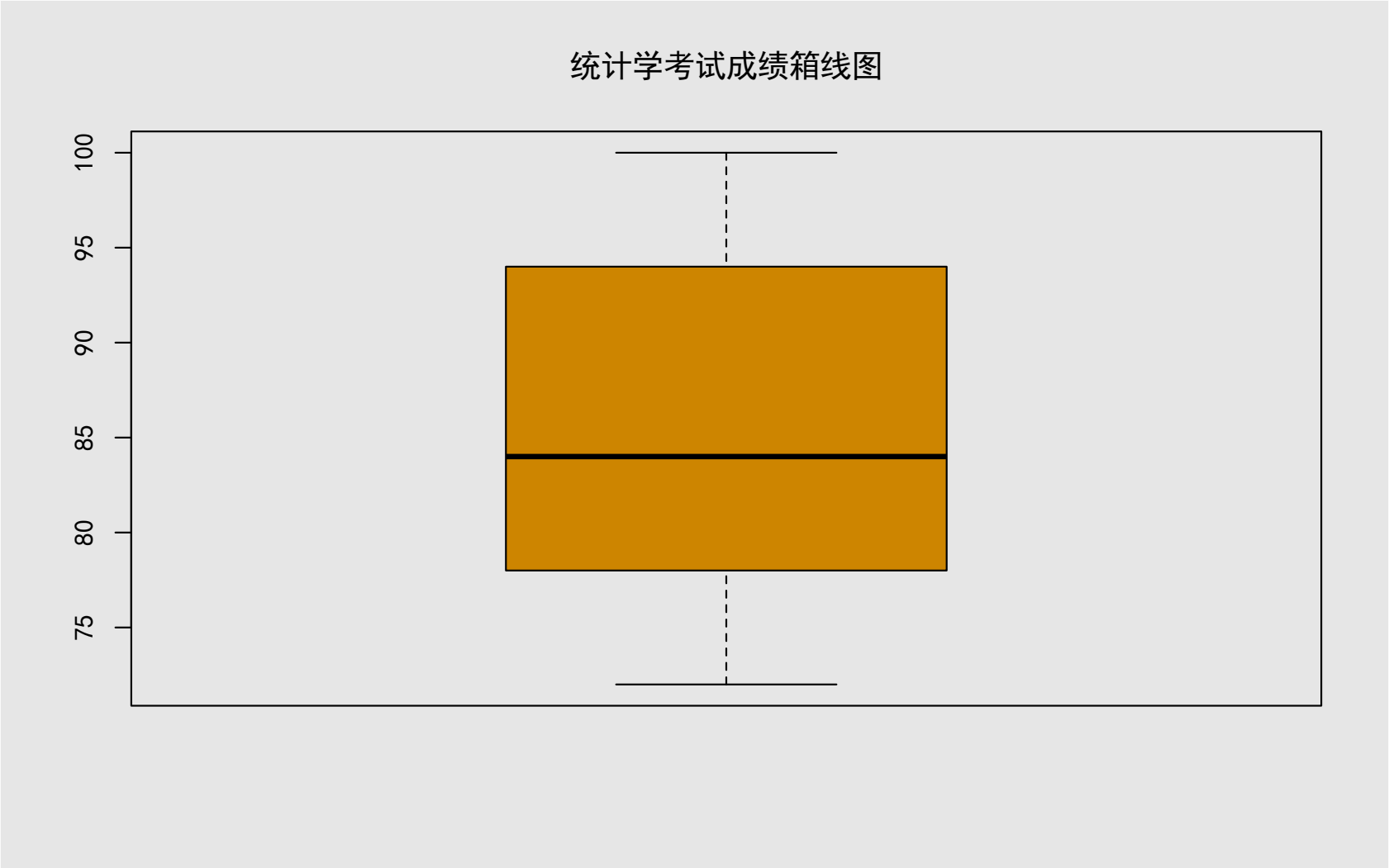
Q2：怎么能大致了解成绩的分布情况?

```
## $breaks
## [1]  70  75  80  85  90  95 100
##
## $counts
## [1]  5 14  7  8  8  8
```

Q3：能否用图示的方法表现上述结论?

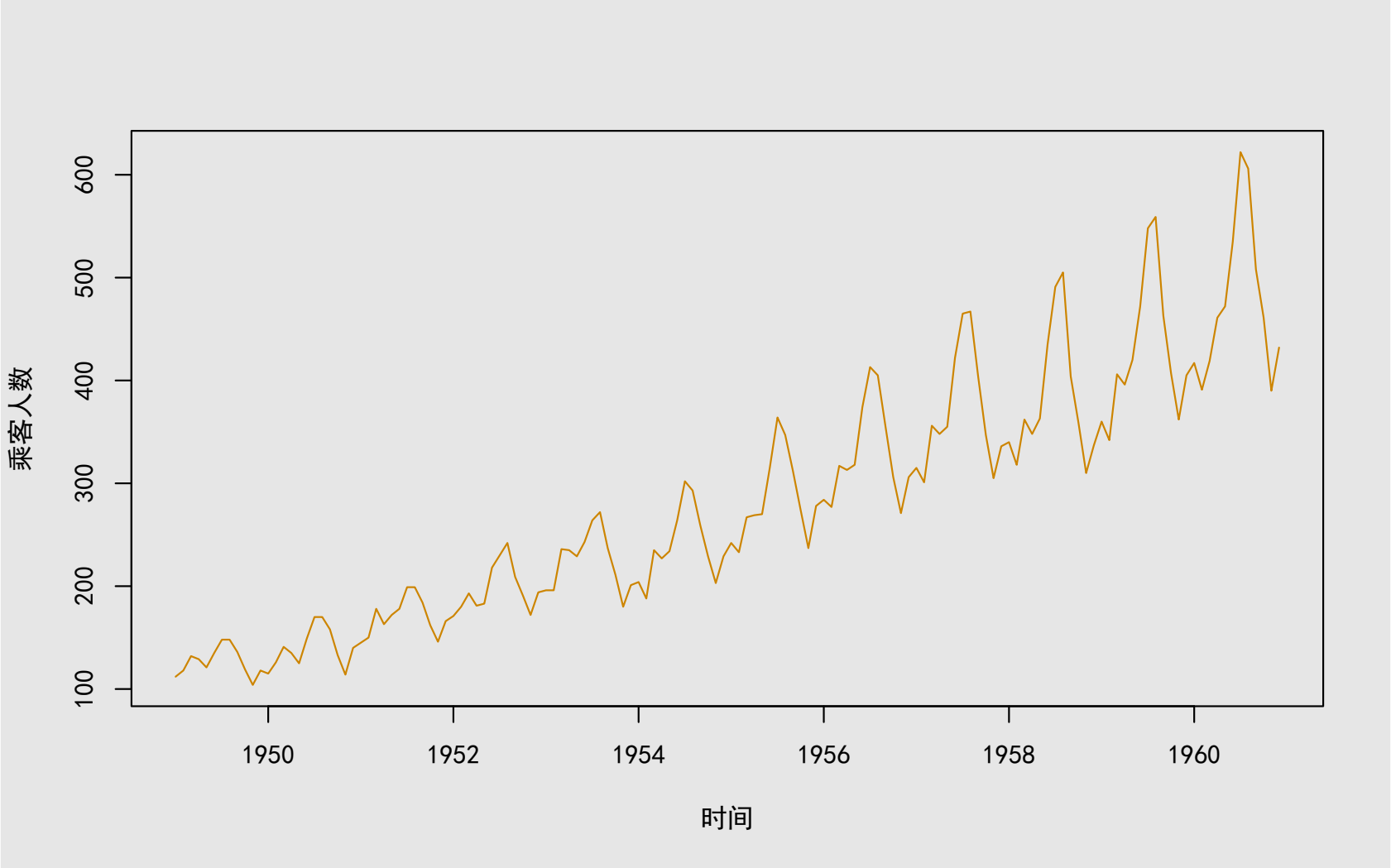


箱线图（Boxplot）



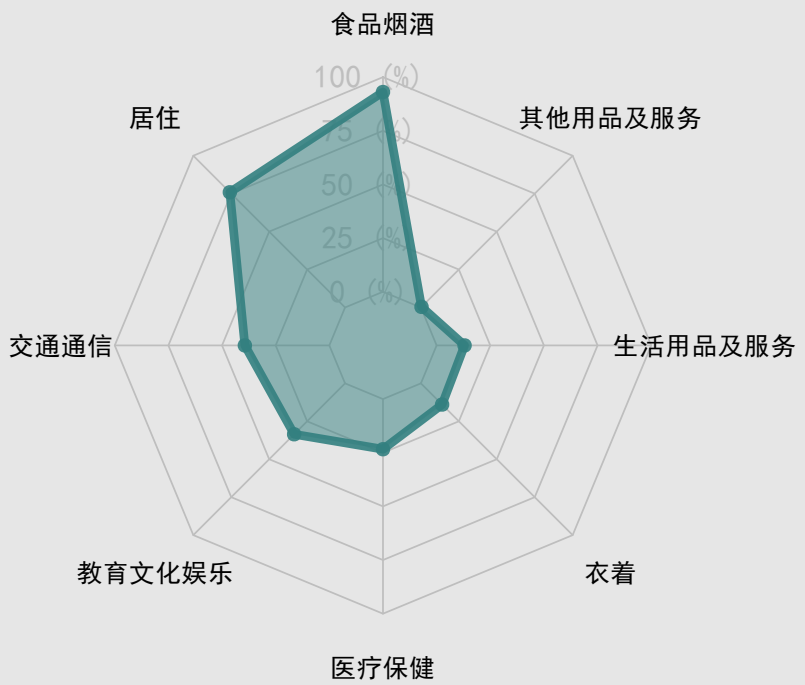
线图（Line chart）

可用于描述事物的动态变化规律



雷达图（Radar chart）

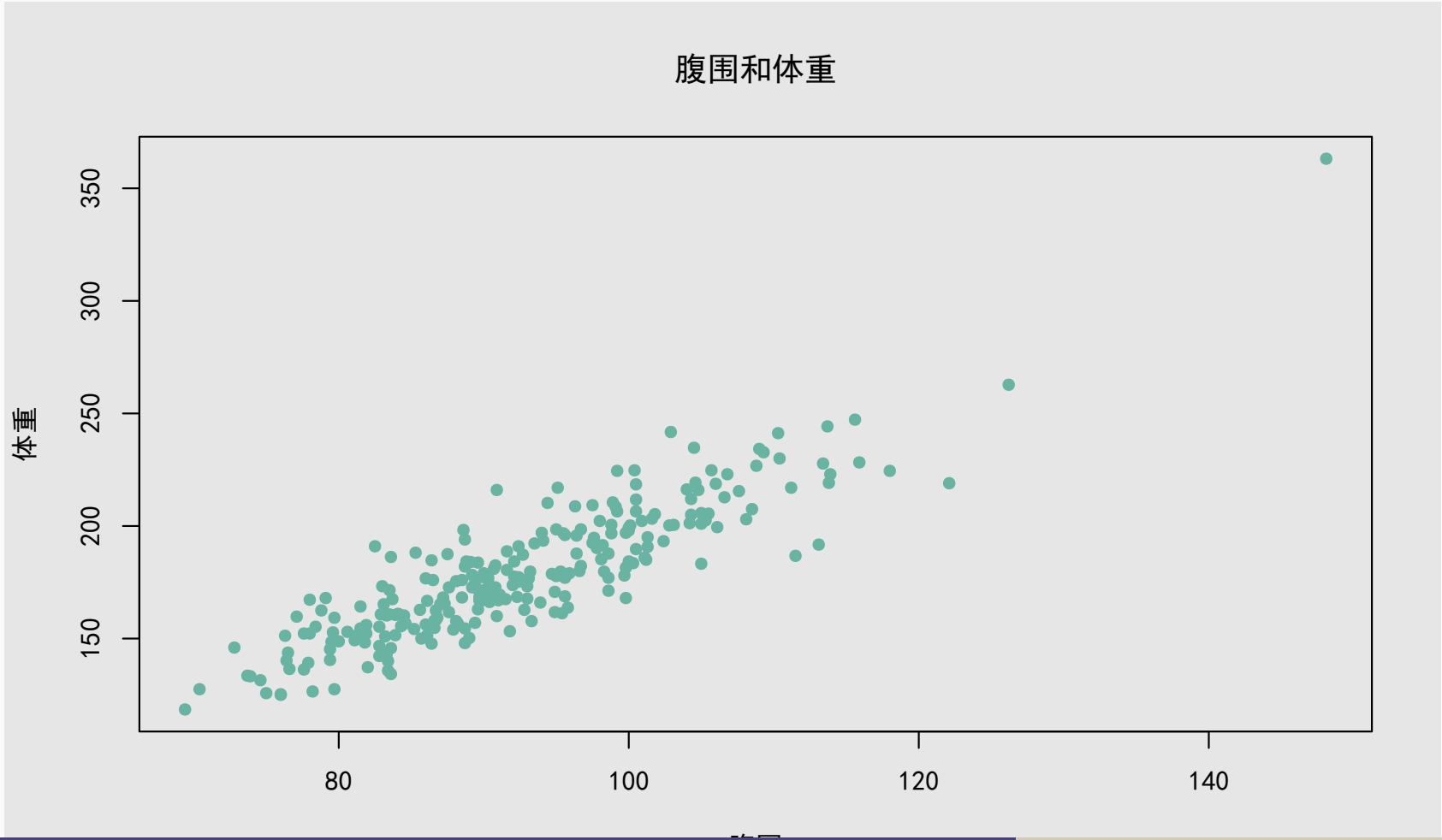
2019 年居民消费支出情况（来源：国家统计局网站）



散点图（Scatter plot）

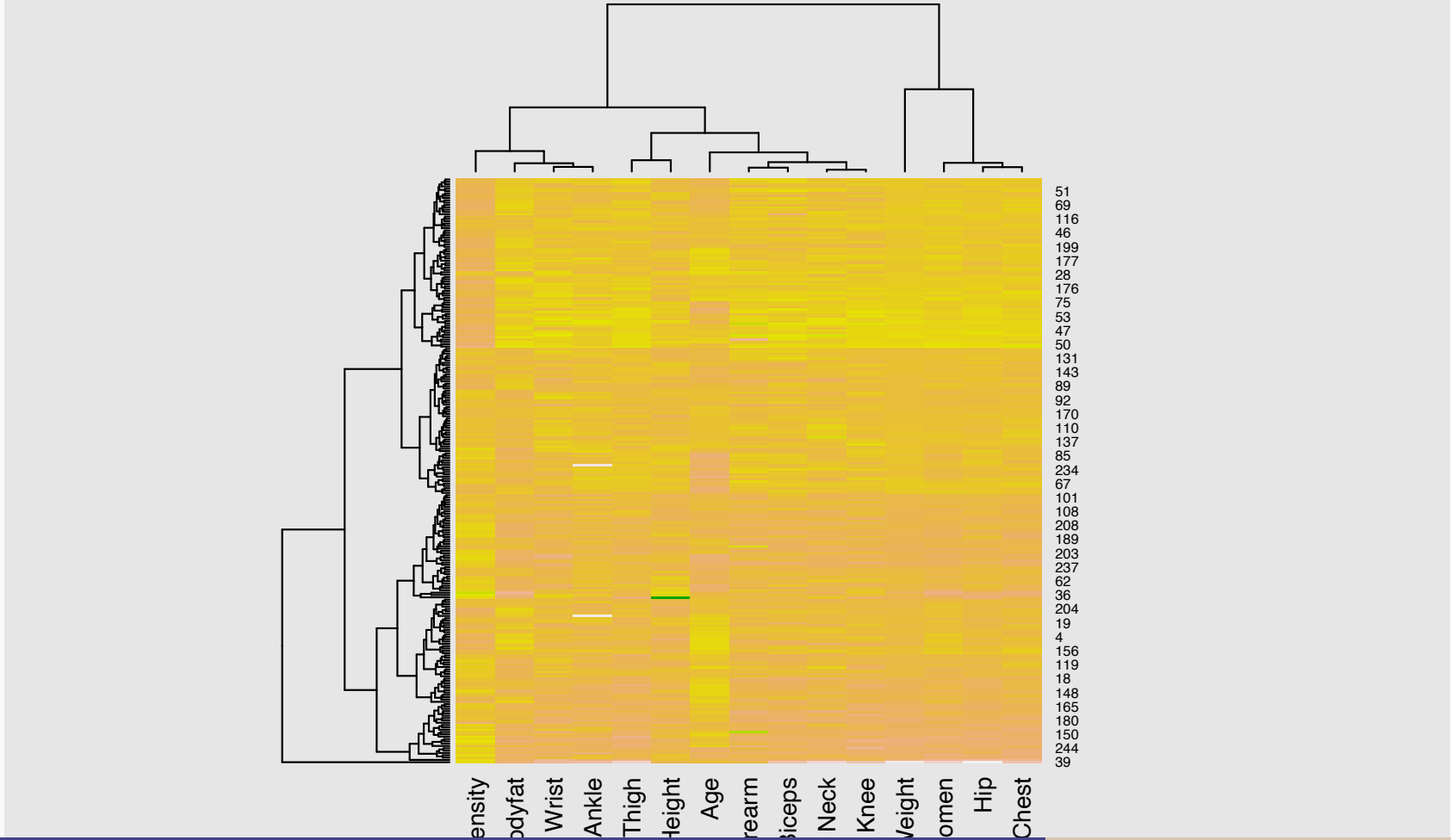
可以反映两个变量之间的相关关系

例：体重与腹围

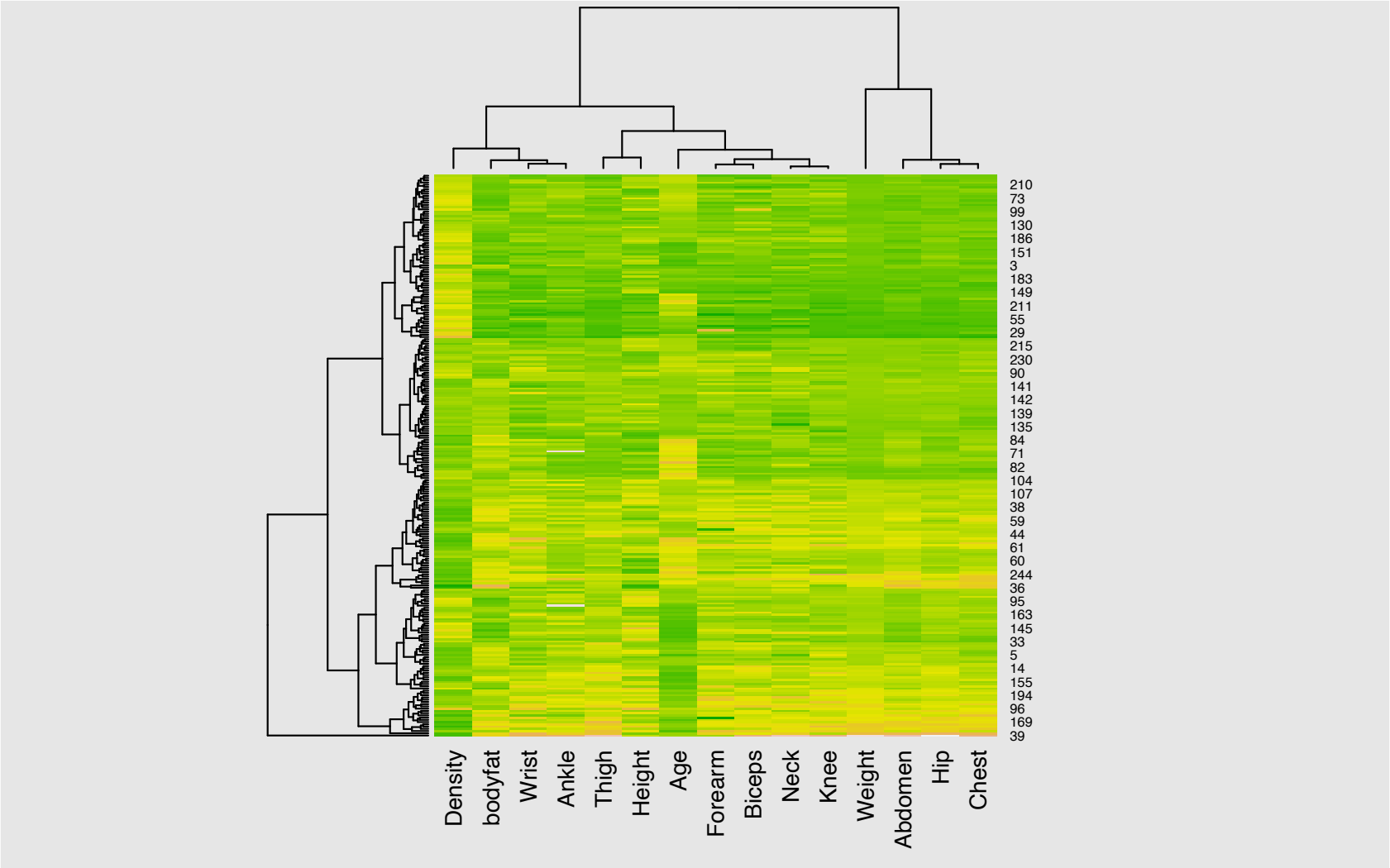


热力图（Heatmap）

- 每一列是一个变量，每一行是一个样本
- 颜色深浅表示取值大小



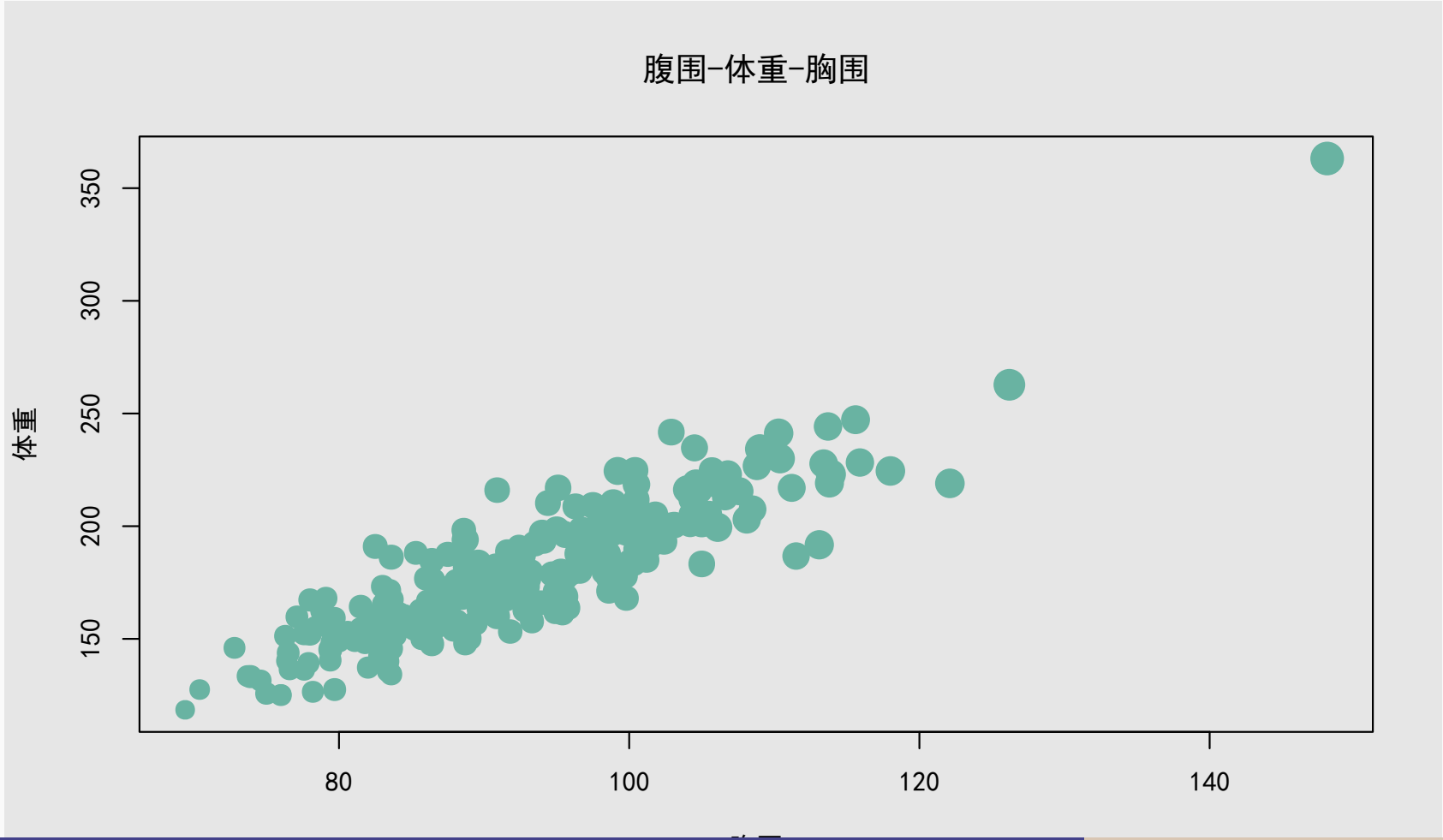
热力图


210
73
99
130
186
151
3
183
149
211
55
29
215
230
90
141
142
139
135
84
71
82
104
107
38
59
44
61
60
244
36
95
163
145
33
5
14
155
194
96
169
39
Density
bodyfat
Wrist
Ankle
Thigh
Height
Age
Forearm
Biceps
Neck
Knee
Weight
Abdomen
Hip
Chest

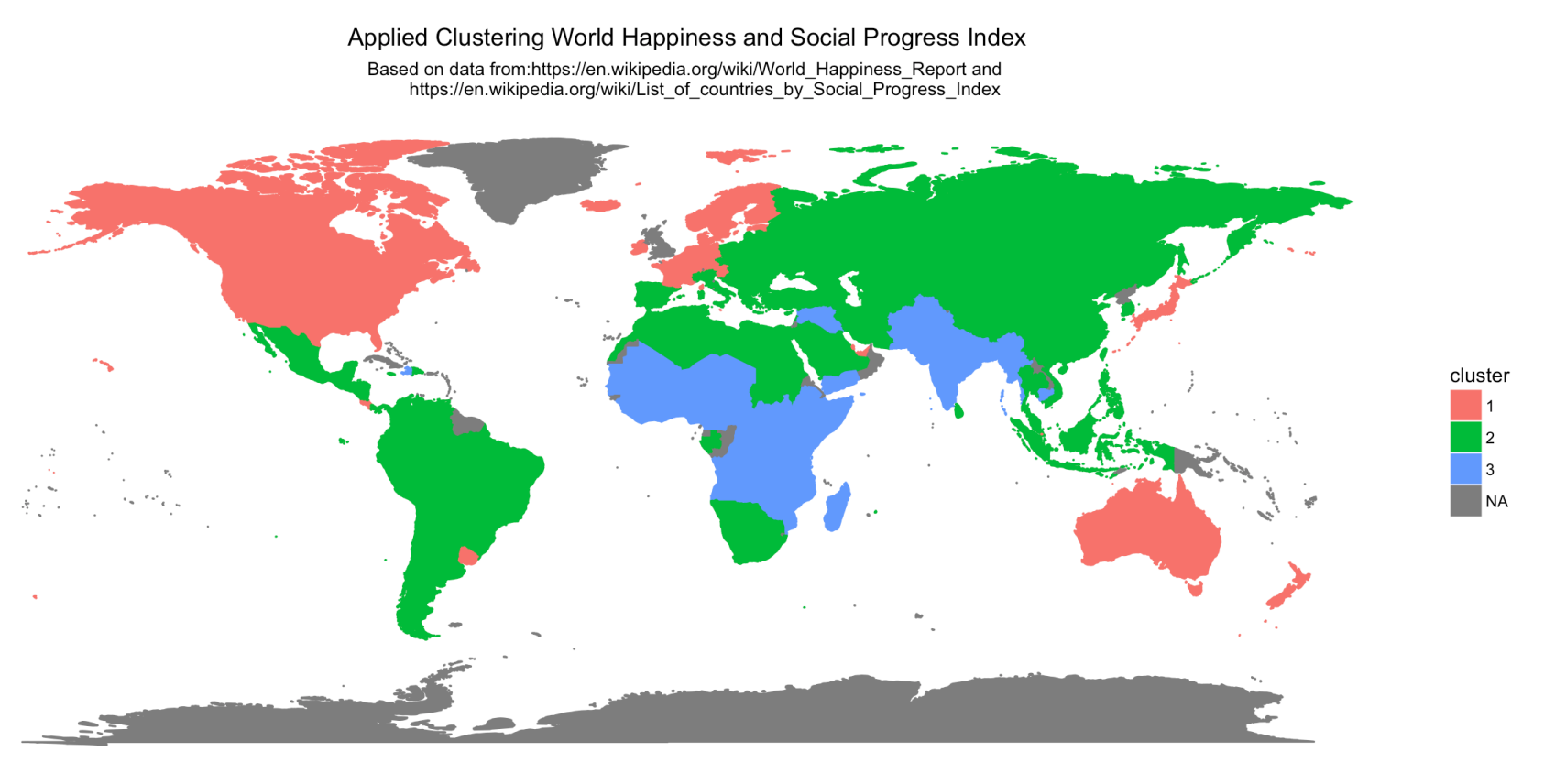
气泡图（Bubble plot）

实质上是散点图

例：体重与腹围，气泡图可以展示三个变量



地图（Map）


Applied Clustering World Happiness and Social Progress Index
Based on data from:https://en.wikipedia.org/wiki/World_Happiness_Report and
https://en.wikipedia.org/wiki/List_of_countries_by_Social_Progress_Index
cluster
1
2
3
NA

描述性统计的基本原则

1. 直观：能清晰地表达主题
2. 醒目：信息点特别突出
3. 变化：图形使用应丰富多彩
4. 图、文并茂：文字应与图、表的篇幅平衡
5. 详略得当，防止冗长的流水帐

描述性统计的分析重点

1. 规律趋势
2. 关联关系
3. 特殊现象
4. 原因分析

《鲜活的数据——数据可视化指南》

统计学，其实是用数据讲故事

- 我们手头的大量数据反映了真实世界。这些故事反过来可以帮助我们解决真实世界中存在的问题。
- 在一堆一堆的数字之间存在着实际的意义、真相和美学。
- 有意思的故事重点：规律、异常（原因分析）
- 当图表都展现在你的眼前时，请问这些结果的意义何在，它们是否在你的意料之中？有没有结果让你感到惊讶?
- 在故事中，观众会感觉到情绪、信念和激情

本讲总结

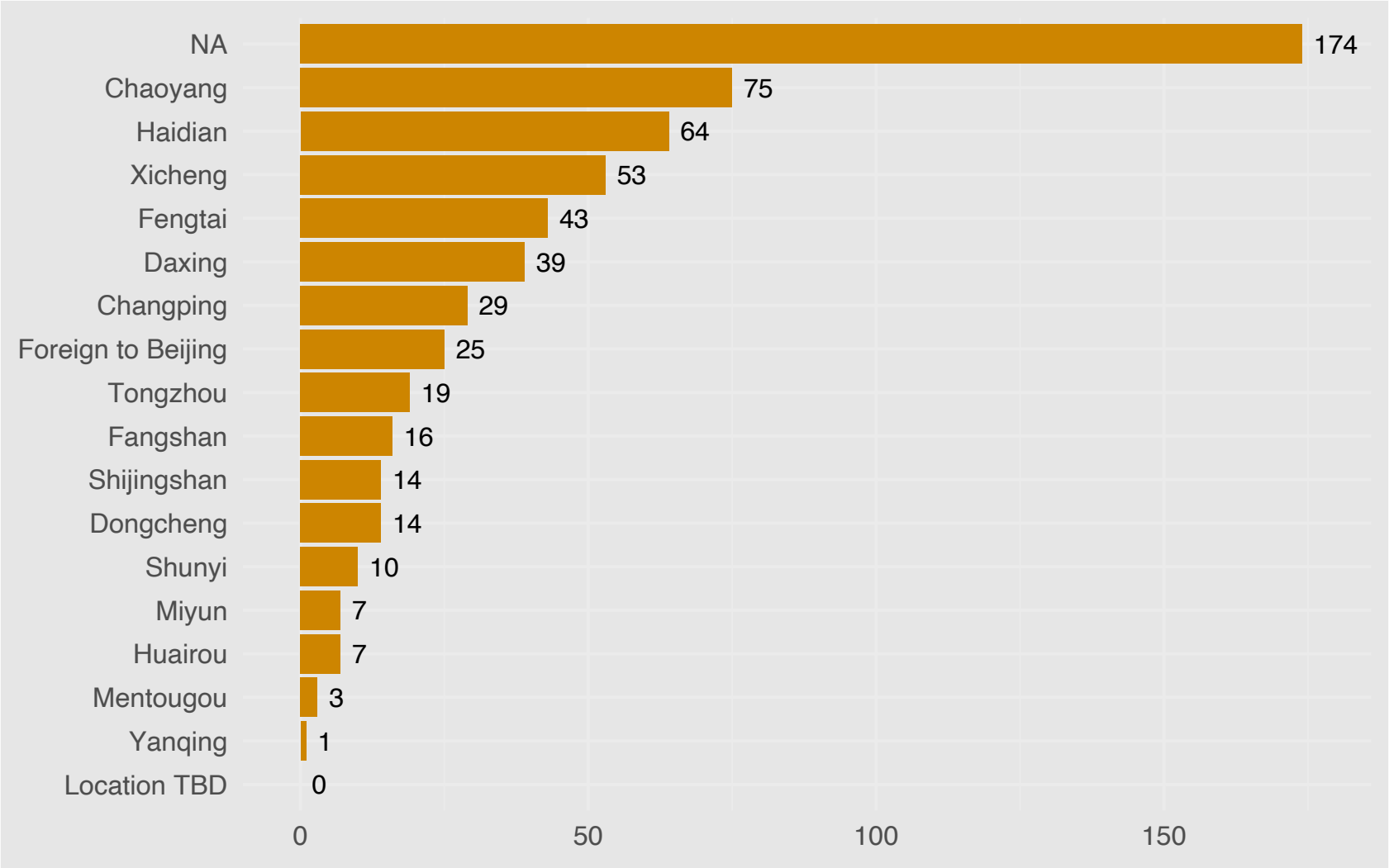
1. 汇总统计量 - 用少量数字来描述数据
2. 统计图 - 用图将数据可视化

Outline

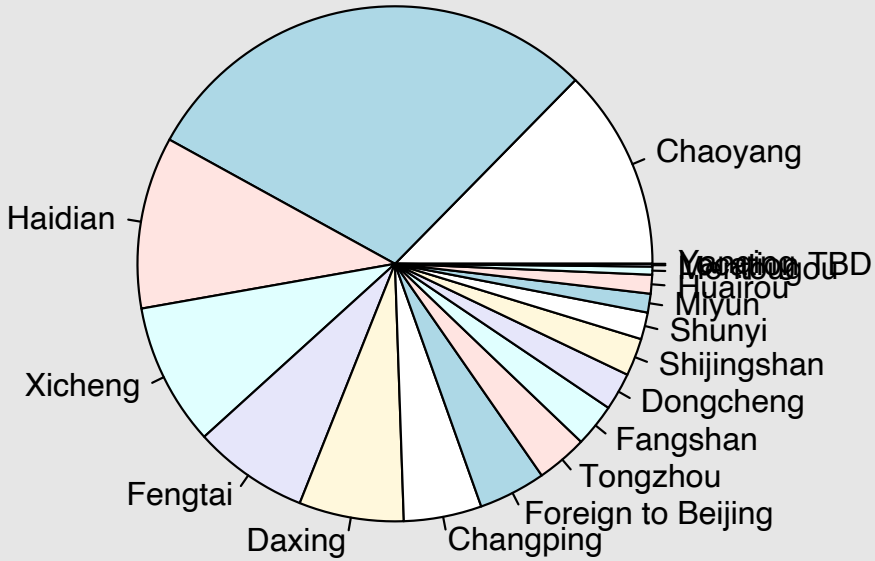
北京市

```
##        name nowConfirm confirm suspect dead de
## 1 Chaoyang         75      75       0    0
## 2      <NA>         68     174       0    0
## 3   Haidian         64      64       0    0
## 4   Xicheng         53      53       0    0
## 5   Fengtai         43      43       0    0
## 6    Daxing         39      39       0    0
##    showHeal
## 1     FALSE
## 2      TRUE
## 3     FALSE
```

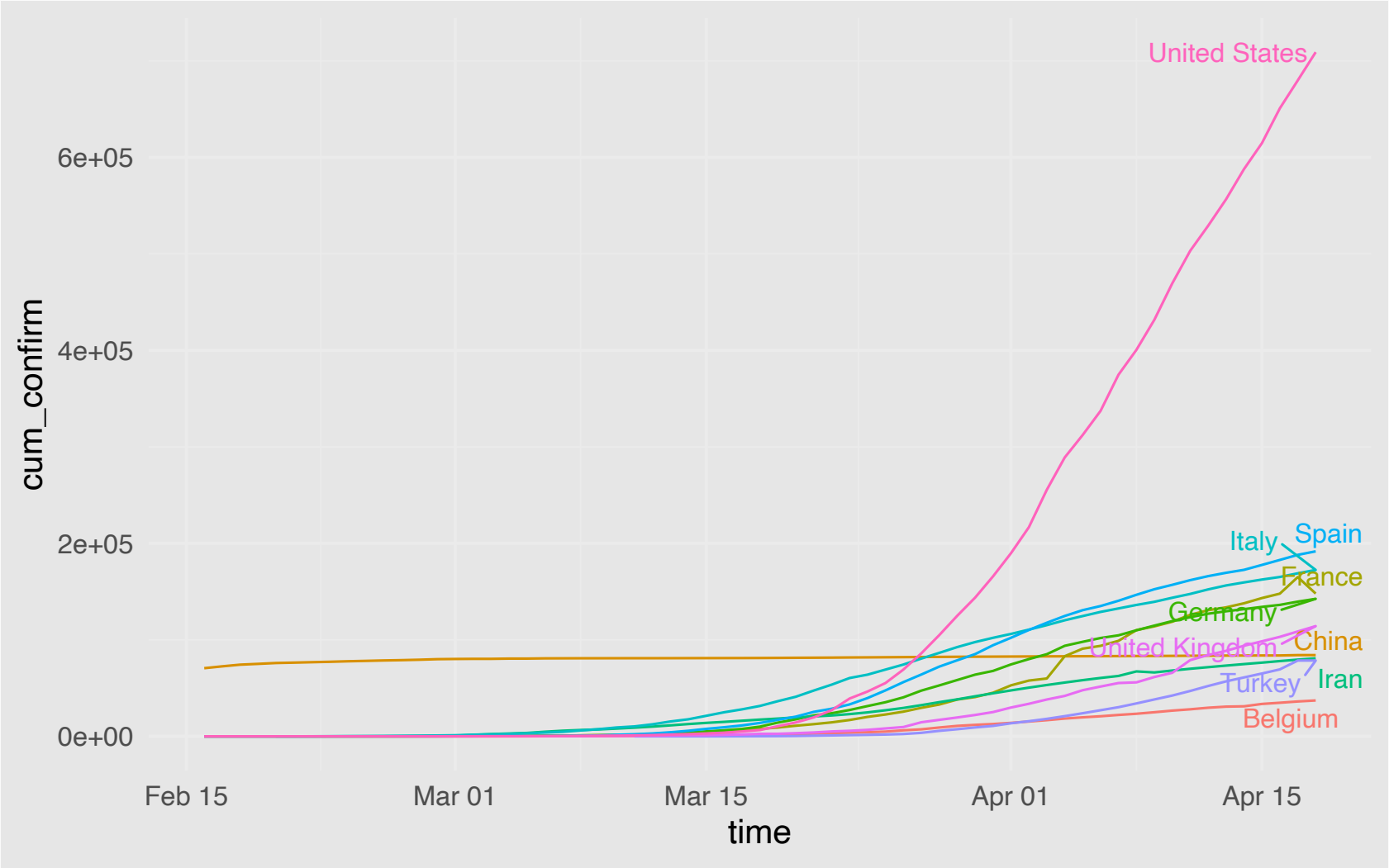
北京每个区的分布?


NA
Chaoyang
Haidian
Xicheng
Fengtai
Daxing
Changping
Foreign to Beijing
Tongzhou
Fangshan
Shijingshan
Dongcheng
Shunyi
Miyun
Huairou
Mentougou
Yanqing
Location TBD
174
75
64
53
43
39
29
25
19
16
14
14
10
7
7
3
1
0
0
50
100
150

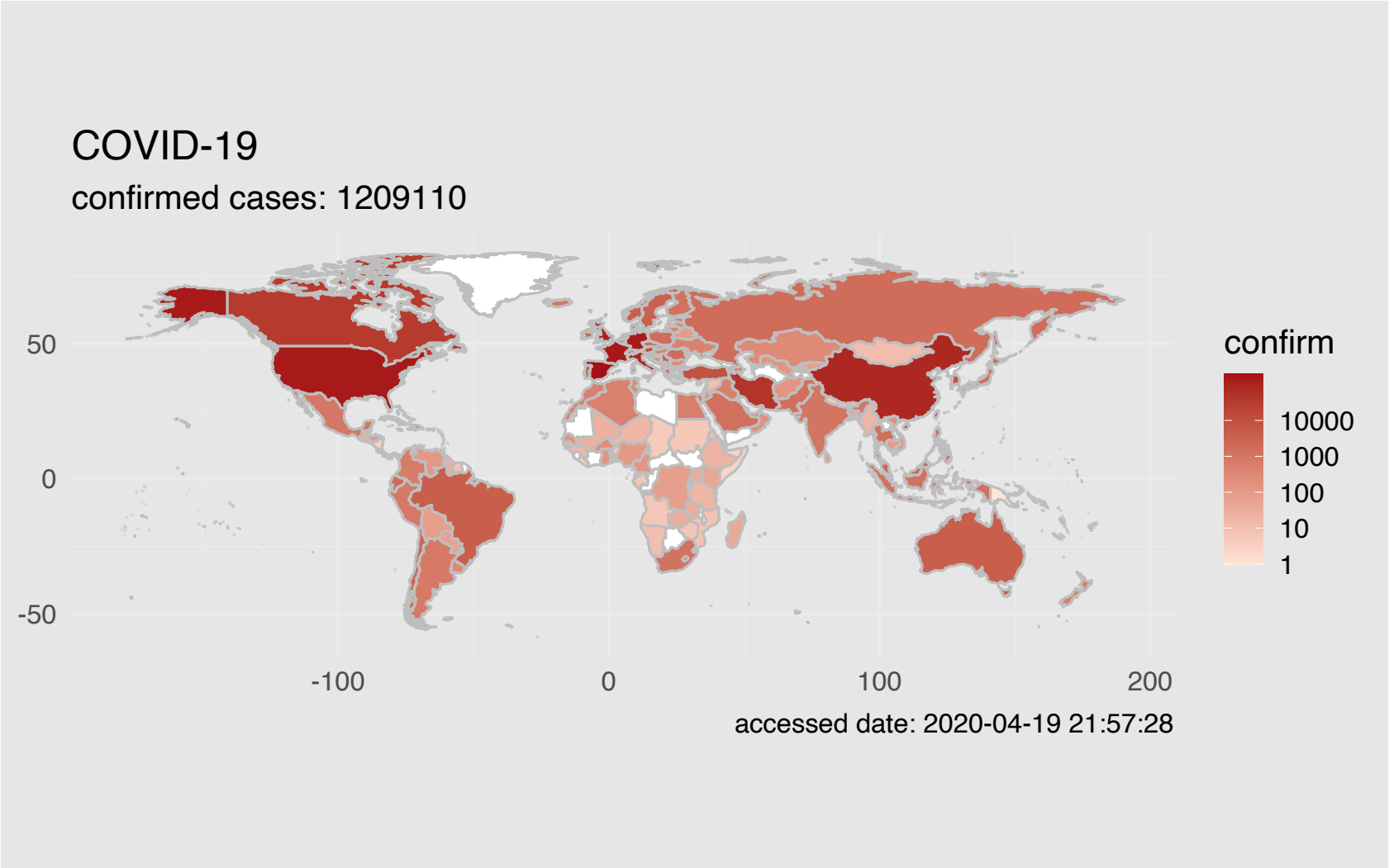
北京每个区的分布?


Chaoyang
Haidian
Xicheng
Fengtai
Daxing
Changping
Foreign to Beijing
Tongzhou
Fangshan
Dongcheng
Shijingshan
Shunyi
Miyun
Huairou

TOP10 国家的变化趋势



地图


COVID-19
confirmed cases: 1209110
50
0
-50
-100
0
100
200
accessed date: 2020-04-19 21:57:28
confirm
10000
1000
100
10
1

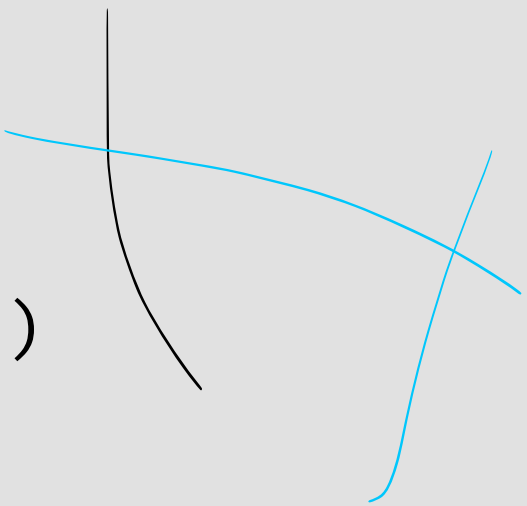
Outline

汇总统计量

- `mean()`
- `median()`
- `table()`
- `var()`
- `sd()`
- `summary()`

例子

```
scores <- c(9.9, 9.3, 9.3, 9.3, 9.2,
            8.9, 8.8, 8.8, 8.7, 8.5, 8.4)
scores
mean(scores)
median(scores)
table(scores)
summary(scores)
var(scores)
sd(scores)
```



统计图

- barplot()
- pie()
- hist()
- boxplot()
- plot()
- fmsb::radarchart()
- heatmap()

例子：条形图

```
浏览器 <- c("Chrome", "Edge", "Firefox", "IE",
            "Opera", "Safari", "Others")
用户数 <- c(2502.4, 150.78, 395.83,
          238.05, 86.49, 387.65, 134.8)
IB <- data.frame(浏览器, 用户数)
barplot(height = IB$用户数,
        names.arg = IB$浏览器)
```

例子：饼图

```
pie(IB$用户数,
    labels = IB$浏览器,
    main = "2018 年浏览器使用情况")
```

例子：气泡图

```
# fat <- read.csv('./data/bodyfat.csv')
plot(fat$Abdomen, fat$Weight,
     cex=fat$Chest/50,
     col="#69b3a2",
     pch=16)
```

例子：箱线图

```
boxplot(weight~feed, data=chickwts)
```

Outline

1. 请收集和分析全班同学的：性别、年龄、身高、体重、籍贯、星座的数据，做汇总统计和描述性统计。（请课代表协助收集数据）
2. （选做）运用 COVID-19 的国际数据，通过描述性统计方法，进行各个国家防控形势的对比分析。

Final words

- COVID-19 相关数据可见诸多 R 包（如 nCov2019: https://github.com/GuangchuangYu/nCov2019）.
- 如果你想学习更多更有意思的可视化，可参考: https://www.r-graph-gallery.com/.
- 也可以参考以下 R 包:
  - ggplot2: https://ggplot2.tidyverse.org/.
  - plotly: https://plotly.com/r/.
  - shiny: https://shiny.rstudio.com/.
- 习惯使用帮助文档。