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

第 4 讲：聚类分析

康雁飞

2020-05-25

Outline

1. 为什么聚类?
2. 如何度量相似性?
3. 如何聚类?
4. 作业

Outline

1 为什么聚类?

2 如何度量相似性?

3 如何聚类?

4 作业

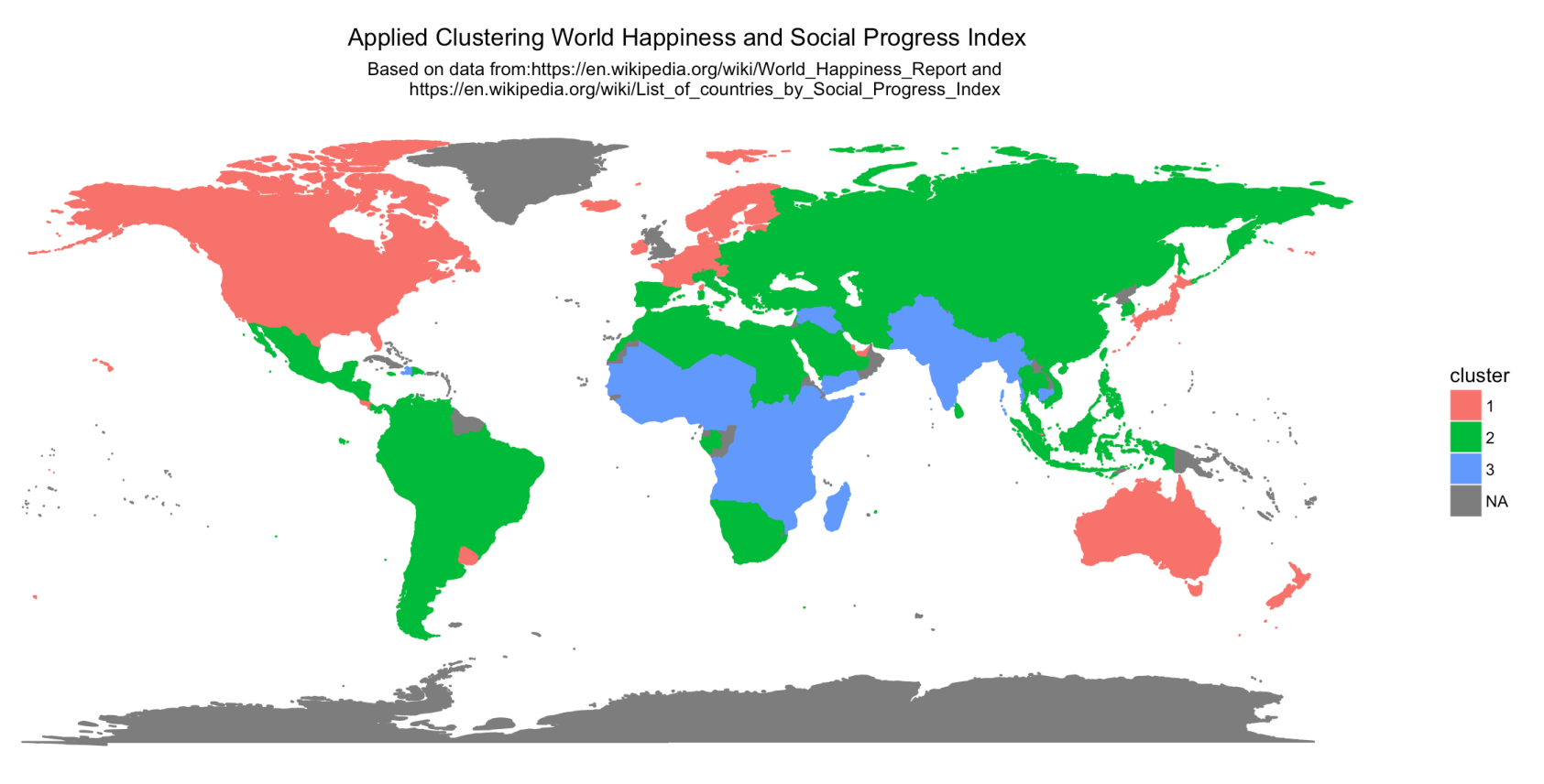
为什么聚类?

1. 海量数据
2. 指标多而杂
3. 数据内在特征不直观

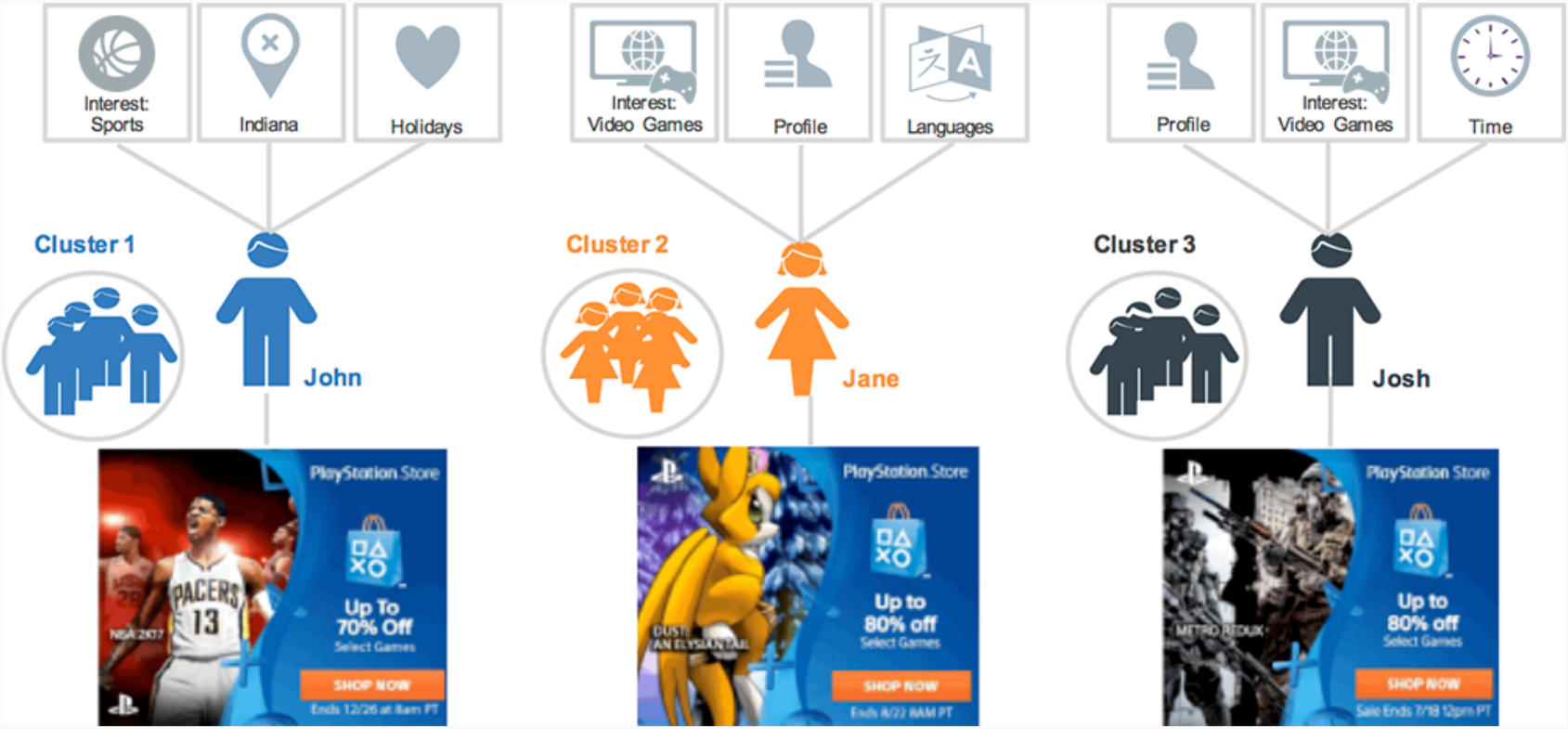
⇓

1. 聚类分析（对数据）
2. 主成分分析（对指标）

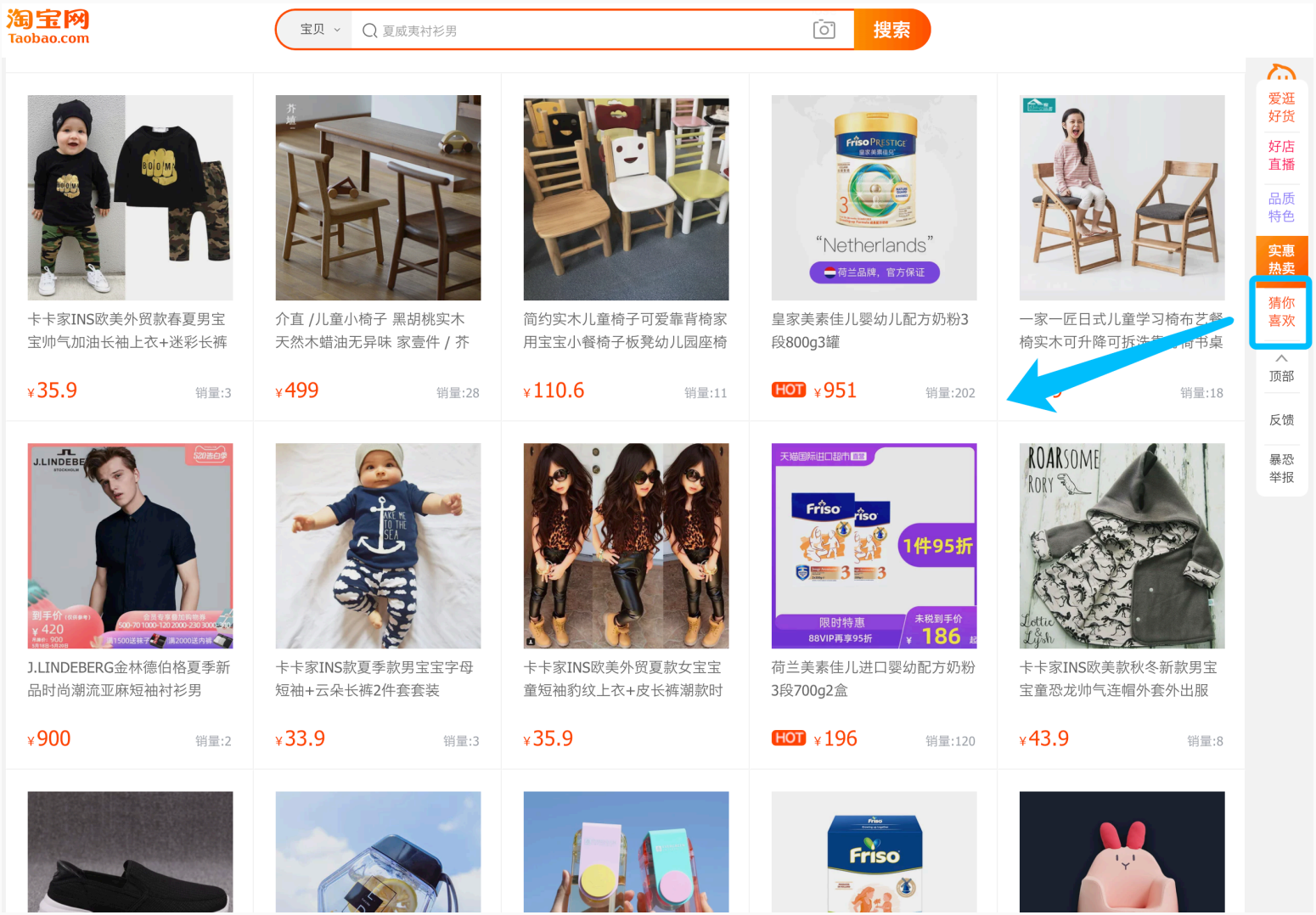
举例：幸福指数


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

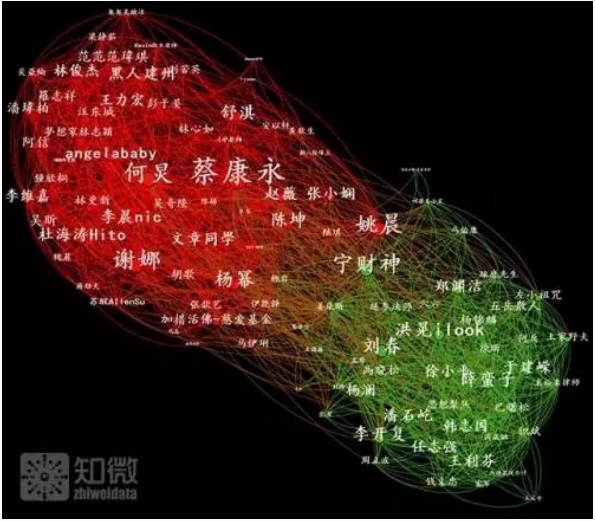
举例：顾客行为


Interest: Sports
Indiana
Holidays
Cluster 1
John
PlayStation Store
Up To 70% Off
Select Games
SHOP NOW
Interest: Video Games
Profile
Languages
Cluster 2
Jane
PlayStation Store
Up to 80% off
Select Games
SHOP NOW
Profile
Interest: Video Games
Time
Cluster 3
Josh
PlayStation Store
Up to 80% off
Select Games
SHOP NOW

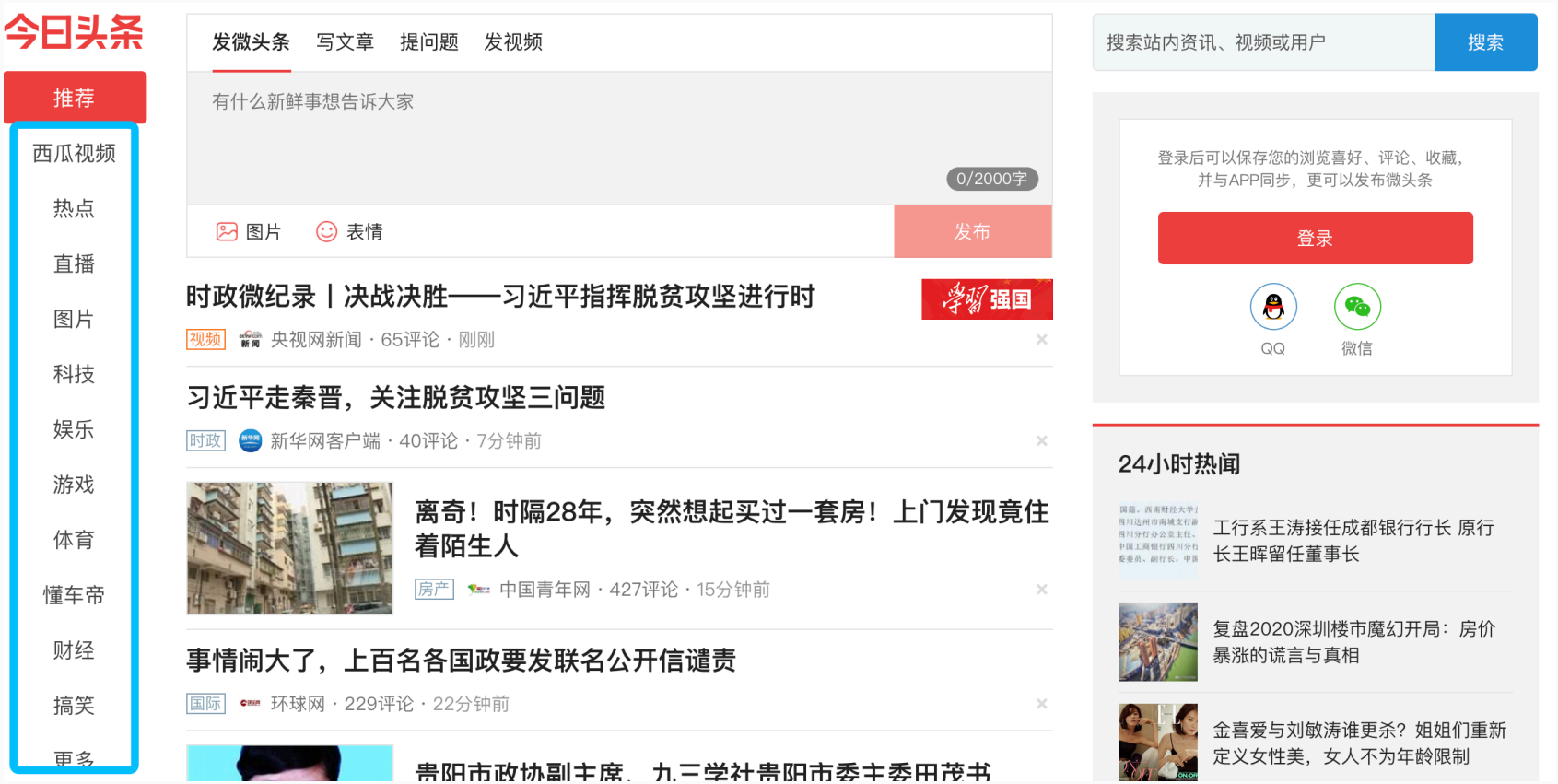
举例：顾客行为


淘宝网
Taobao.com
宝贝
夏威夷衬衫男
搜索
卡卡家INS欧美外贸款春夏男宝宝帅气加油长袖上衣+迷彩长裤
¥35.9
销量:3
介直 /儿童小椅子 黑胡桃实木 天然木蜡油无异味 家壹件 / 芥
¥499
销量:28
简约实木儿童椅子可爱靠背椅家用宝宝小餐椅子板凳幼儿园座椅
¥110.6
销量:11
皇家美素佳儿婴幼儿配方奶粉3段800g3罐
HOT ¥951
销量:202
一家一匠日式儿童学习椅布艺餐椅实木可升降可拆洗
销量:18
爱逛好货
好店直播
品质特色
实惠热卖
猜你喜欢
顶部
反馈
暴恐举报
J.LINDEBERG金林德伯格夏季新品时尚潮流亚麻短袖衬衫男
¥900
销量:2
卡卡家INS款夏季款男宝宝字母短袖+云朵长裤2件套套装
¥33.9
销量:3
卡卡家INS欧美外贸夏款女宝宝童短袖豹纹上衣+皮长裤潮款时
¥35.9
荷兰美素佳儿进口婴幼配方奶粉3段700g2盒
HOT ¥196
销量:120
卡卡家INS欧美款秋冬新款男宝宝童恐龙帅气连帽外套外出服
¥43.9
销量:8

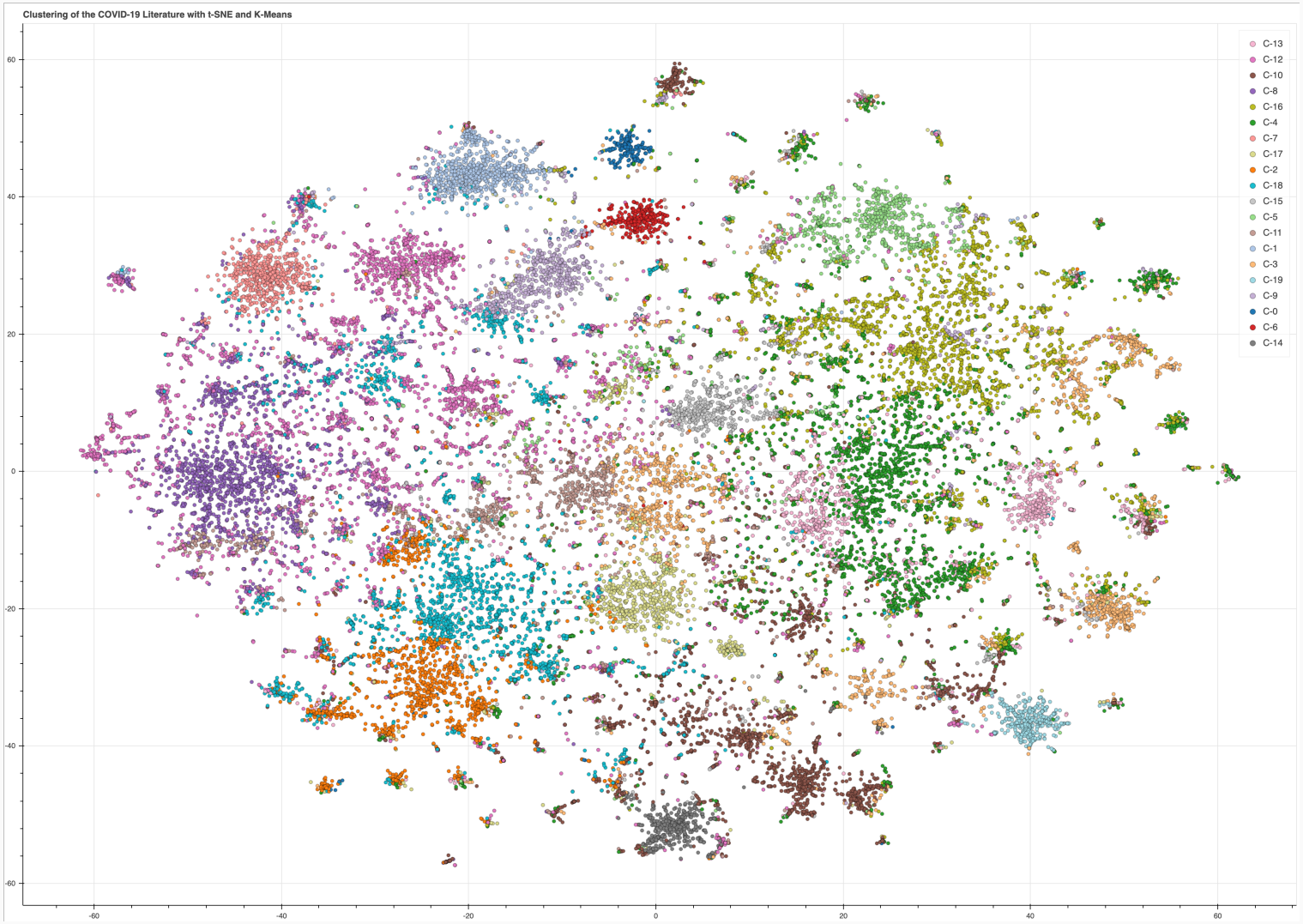
举例：社交网络


范范范玮琪
林俊杰
黑人建州
罗志祥
王力宏
潘玮柏
汪东城
舒淇
林心如
阿信
angelababy
何炅
蔡康永
赵薇
张小娴
李维嘉
李晨nic
陈坤
姚晨
吴昕
杜海涛Hito
文章同学
谢娜
杨幂
宁财神
郑渊洁
左小祖咒
五岳散人
加措活佛-慈爱基金
洪晃ilook
刘春
徐小平
薛蛮子
于建嵘
杨澜
潘石屹
李开复
任志强
王利芬
知微
zhiweidata

举例：文本聚类


今日头条
推荐
西瓜视频
热点
直播
图片
科技
娱乐
游戏
体育
懂车帝
财经
搞笑
更多
发微头条
写文章
提问题
发视频
有什么新鲜事想告诉大家
0/2000字
图片
表情
发布
时政微纪录丨决战决胜——习近平指挥脱贫攻坚进行时
学习强国
视频
央视网新闻 · 65评论 · 刚刚
习近平走秦晋，关注脱贫攻坚三问题
时政
新华网客户端 · 40评论 · 7分钟前
离奇！时隔28年，突然想起买过一套房！上门发现竟住着陌生人
房产
中国青年网 · 427评论 · 15分钟前
事情闹大了，上百名各国政要发联名公开信谴责
国际
环球网 · 229评论 · 22分钟前
搜索站内资讯、视频或用户
搜索
登录后可以保存您的浏览喜好、评论、收藏，并与APP同步，更可以发布微头条
登录
QQ
微信
24小时热闻
工行系王涛接任成都银行行长 原行长王晖留任董事长
复盘2020深圳楼市魔幻开局：房价暴涨的谎言与真相
金喜爱与刘敏涛谁更杀？姐姐们重新定义女性美，女人不为年龄限制

举例：COVID19 文献聚类



source: https://maksimekin.github.io/COVID19-Literature-Clustering/plots/t-sne_covid-19_interactive.html

Outline

1 为什么聚类?

2 **如何度量相似性?**

3 如何聚类?

4 作业

相似性度量 - 距离

1. 在聚类分析中，如果样本点为有限维定量指标，常用明考夫斯基距离（Minkowski distance）。
2. 余弦距离（衡量文本之间的相似度通常用）。

明考夫斯基距离（Minkowski distance）

明考夫斯基距离：

$$d_q(x,y)=\left[\sum_{j=1}^{p}|x_j-y_j|^q\right]^{1/q}.$$

1. $q=1$ 时，即绝对值距离：$d_1(x,y)=\sum_{j=1}^{p}|x_j-y_j|$.
2. $q=2$ 时，即欧式距离：$d_2(x,y)=\sqrt{\sum_{j=1}^{p}(x_j-y_j)^2}$.
3. $q=\infty$ 时，即切比雪夫距离：
   $d_\infty(x,y)=\max_{1\le j\le p}|x_j-y_j|$.

明考夫斯基距离

1. 最常用的是欧氏距离。它的优点是：坐标经旋转变换后，点和点之间距离保持不变。
2. 采用明氏距离时，应采用相同量纲的变量。
3. 尽可能避免数据的多重相关性。

余弦距离

余弦距离：

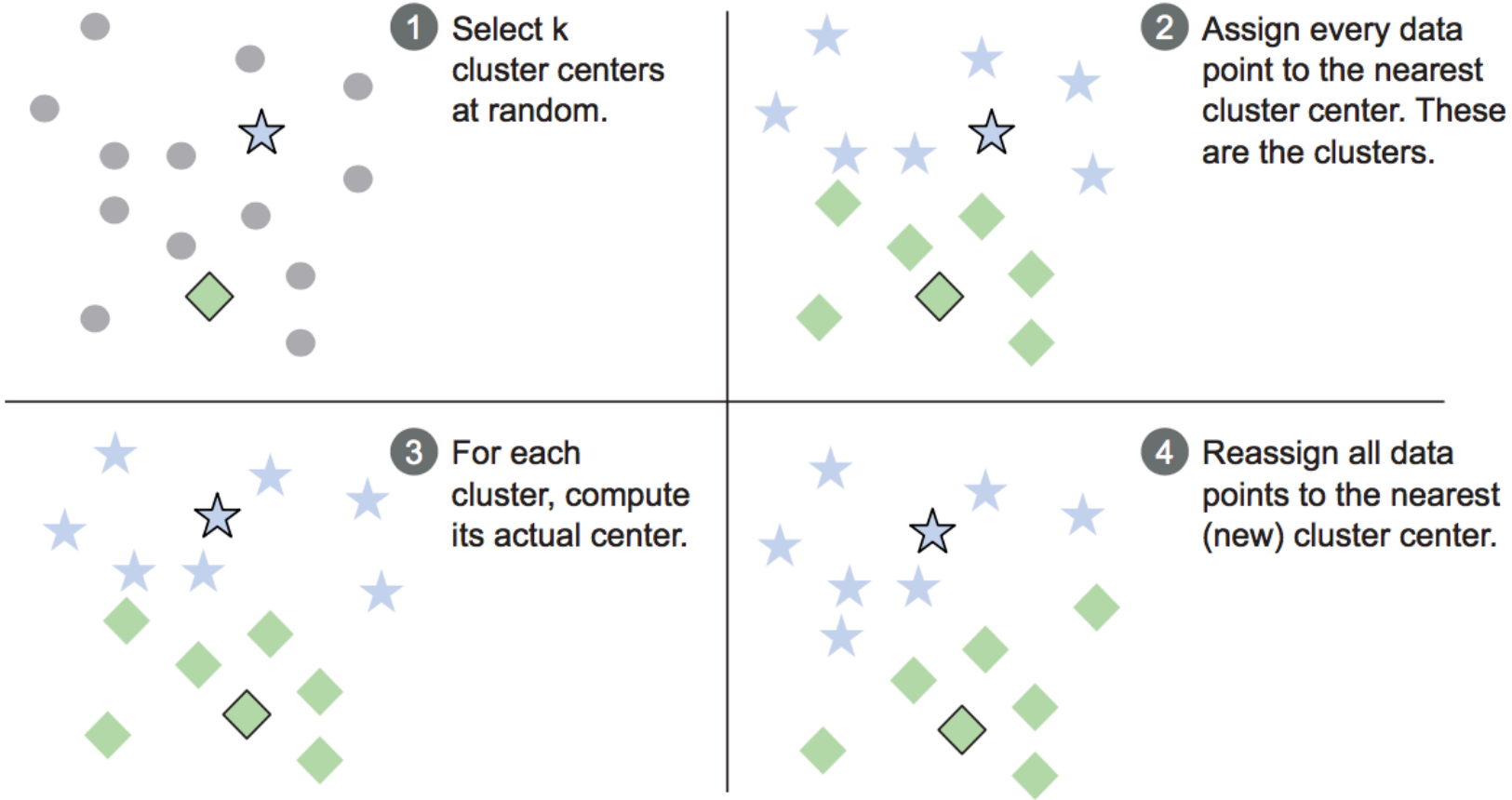
$$cos(x, y) = \frac{x \cdot y}{|x| \cdot |y|}.$$

余弦距离可以衡量不同量纲的变量之间的距离（距离：文本）。

Outline

k-均值聚类（*k*-means clustering）

k-均值聚类是一个迭代聚类算法。



k-均值聚类（k-means clustering）

k-均值聚类

1. 确定类的个数 k。
2. 随机抽取 k 个样本点作为 k 个类中心。
3. 将每个样本点分配到距离其最近的类中。
4. 重新计算类中心。

重复第 3、4 步直到样本点的类别不再变化或者达到了最大迭代次数。

如何选择 k?

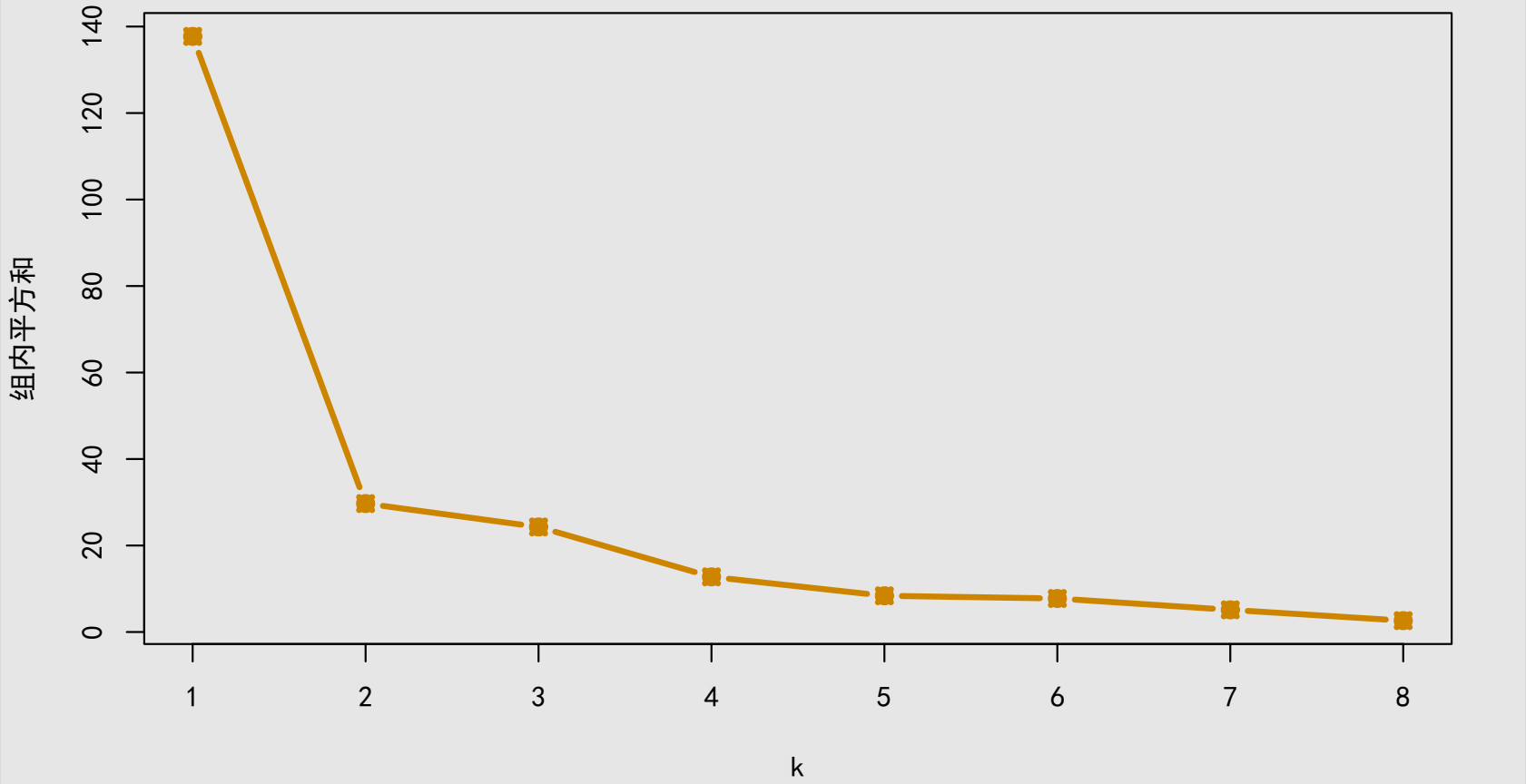
- 如何衡量聚类质量（goodness of clustering)?
- 组内距越小越好，组间距越大越好。
- 组内平方和:

$$tot.within = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2,$$

  其中 μ_k 为第 k 个类 C_k 的中心。
- 问题：组内平方和如何随 k 变化?

如何选择 k?

碎石图（Scree plot）



R 中进行 k-均值聚类

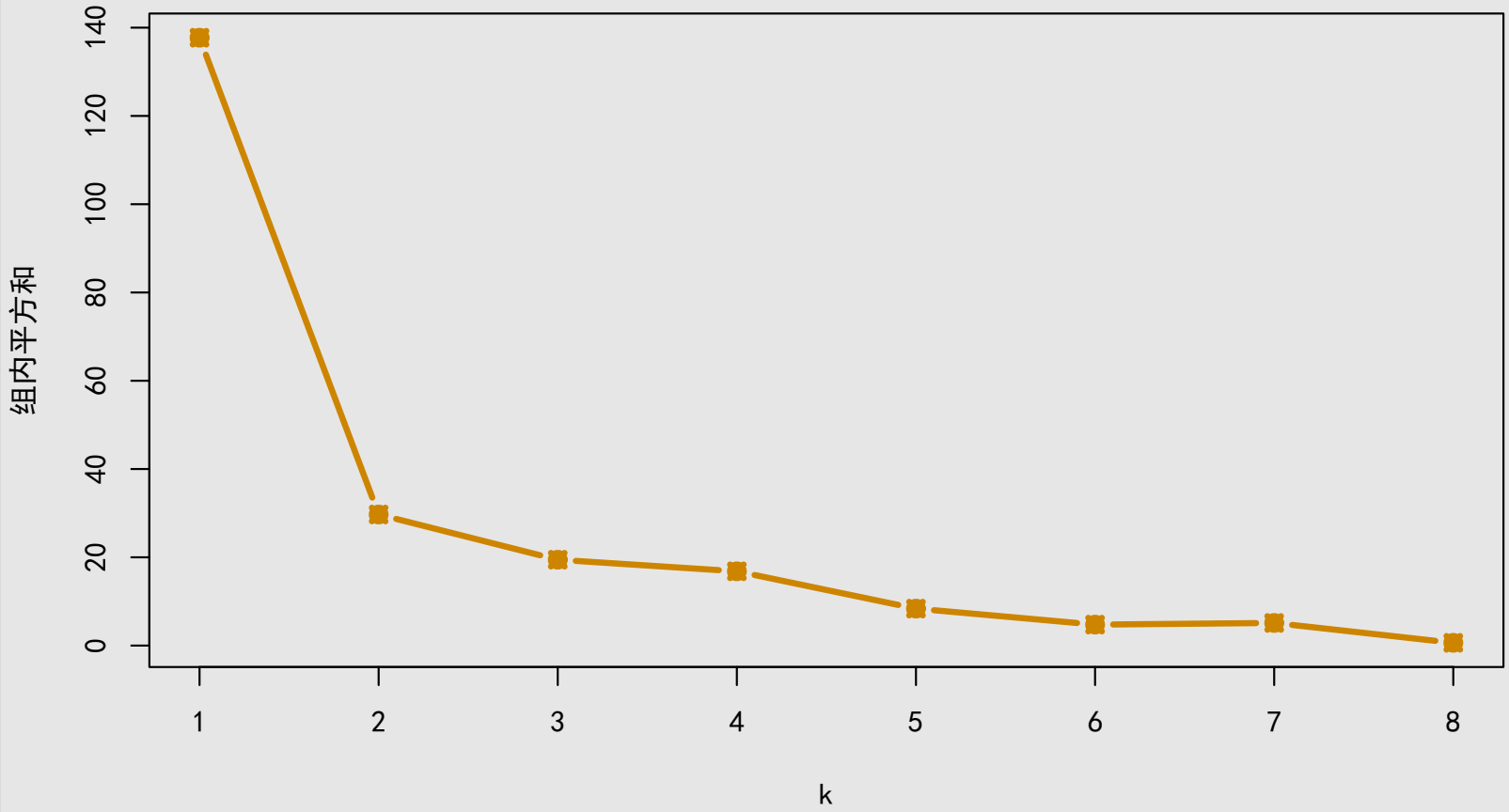
例：某企业生产新式大衣，将新产品的样品寄给九个城市的进货员，并附寄调查意见表征求对新产品的评价，评价分质量、款式、颜色三个方面，以十分评分。

```
coat.scores <- read.csv('./data/coatscores.csv')
str(coat.scores)
```

```
## 'data.frame':    9 obs. of  3 variables:
##  $ Quality: int  3 4 10 8 7 3 8 6 9
##  $ Style  : num  5 2 7.5 9.5 8 4 6.5 3 8.5
##  $ Color  : num  4 5 8.5 7 9 3.5 7 5.5 6
```

选择 k

碎石图（Scree plot）

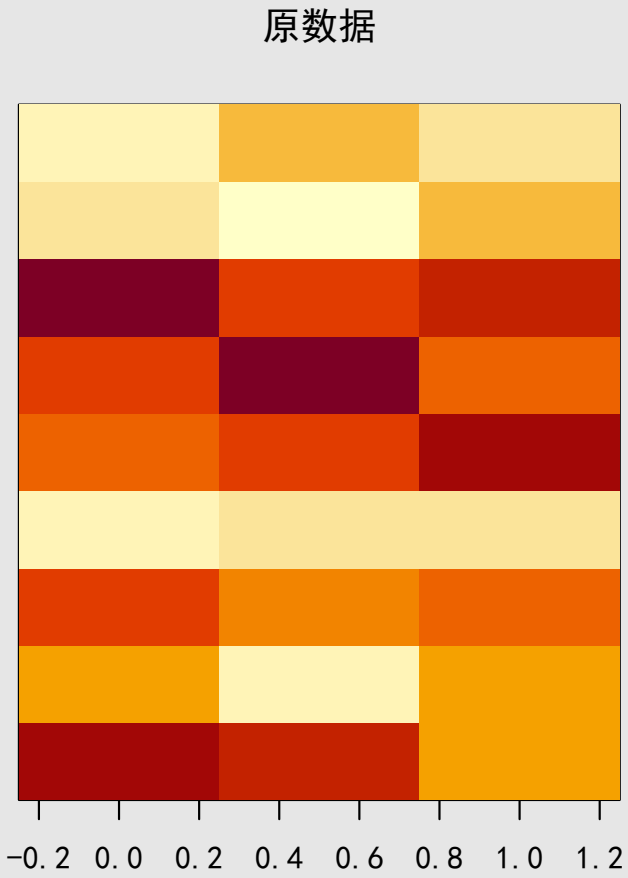


R 中进行 k-均值聚类

```
coat.scores.kmeans <- kmeans(coat.scores, 2)
coat.scores.kmeans
```

```
## K-means clustering with 2 clusters of sizes 4, 5
##
## Cluster means:
##   Quality Style Color
## 1     4.0   3.5   4.5
## 2     8.4   8.0   7.5
##
## Clustering vector:
## [1] 1 1 2 2 2 1 2 1 2
##
## Within cluster sum of squares by cluster:
## [1] 13.5 16.2
##  (between_SS / total_SS =  78.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
```

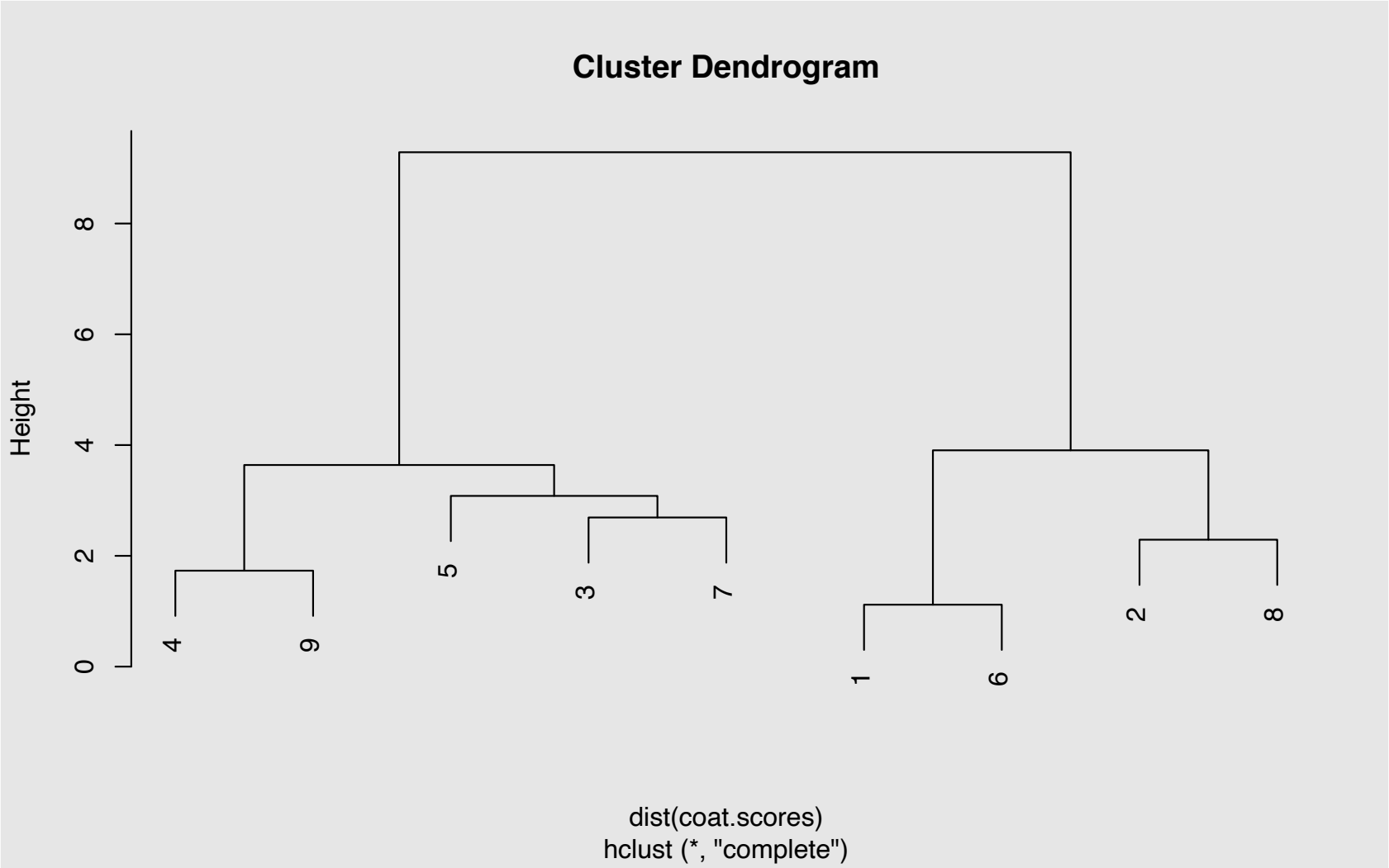
可视化



分层聚类（Hierachical clustering）

1. 分层聚类算法也是一种迭代算法。
2. 它不需要我们提前指定分类个数。
3. 它可以输出一个树状结构。

分层聚类


Cluster Dendrogram
Height
0
2
4
6
8
4
9
5
3
7
1
6
2
8
dist(coat.scores)
hclust (*, "complete")

分层聚类

分层聚类

1. 每个样本点自成一类。
2. 选择最近的两个类聚成一类。
3. 计算新的类与类之间的距离。

重复第 2、3 步直至所有的样本点聚为一类。

如何计算类 C_1 与类 C_2 之间的距离？聚合指数

1. 最短距离法：$D\left(C_1, C_2\right)=\min\limits_{\substack{x_i \in C_1 \\ y_j \in C_2}}\left\{d\left(x_i, y_j\right)\right\}$.
2. 最长距离法：$D\left(C_1, C_2\right)=\max\limits_{\substack{x_i \in C_1 \\ y_j \in C_2}}\left\{d\left(x_i, y_j\right)\right\}$.
3. 重心法：$D\left(C_1, C_2\right)=d(\bar{x}, \bar{y})$.
4. 类平均法：$D\left(C_1, C_2\right)=\frac{1}{l \times m} \sum\limits_{x_i \in C_1} \sum\limits_{y_j \in C_2} d(x_i, y_j)$.

例：大衣评分

首先需要明确：

1. 如何测度距离 $d(x_i, y_j)$？（欧式距离）
2. 如何测度聚合指数 $D(C_1, C_2)$？（最长距离法）

例：大衣评分

```
dist(coat.scores)
```

```
##          1        2        3        4        5        6        7        8
## 2 3.316625
## 3 8.689074 8.860023
## 4 7.365460 8.732125 3.201562
## 5 7.071068 7.810250 3.082207 2.692582
## 6 1.118034 2.692582 9.287088 8.215838 7.889867
## 7 6.020797 6.344289 2.692582 3.000000 2.692582 6.595453
## 8 3.905125 2.291288 6.726812 6.964194 6.184658 3.741657 4.301163
## 9 7.228416 8.261356 2.872281 1.732051 3.640055 7.905694 2.449490 6.284903
```

首先 **1** 和 **6** 合并为一类。

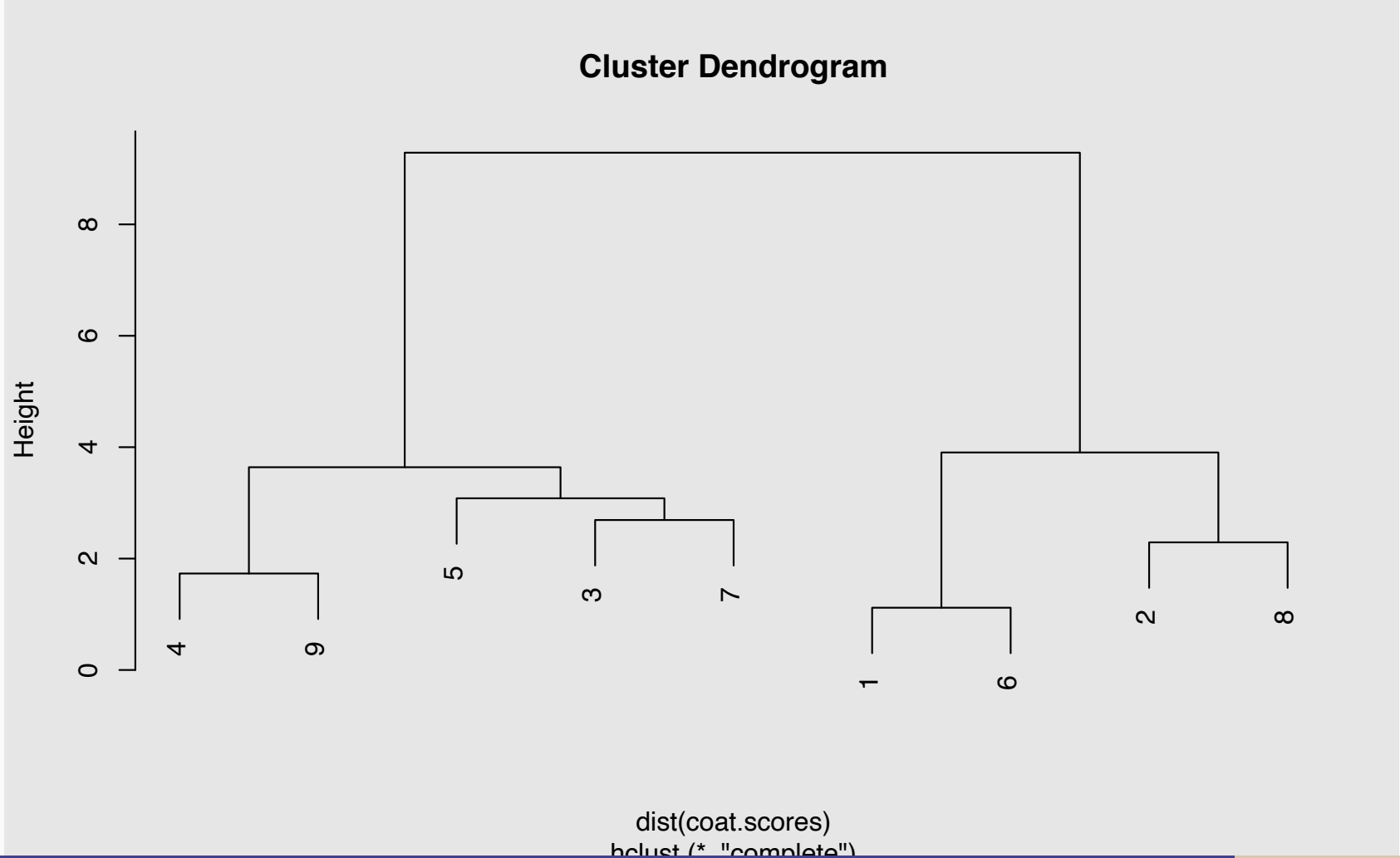
例：大衣评分

```
coat.scores.hc$merge
```

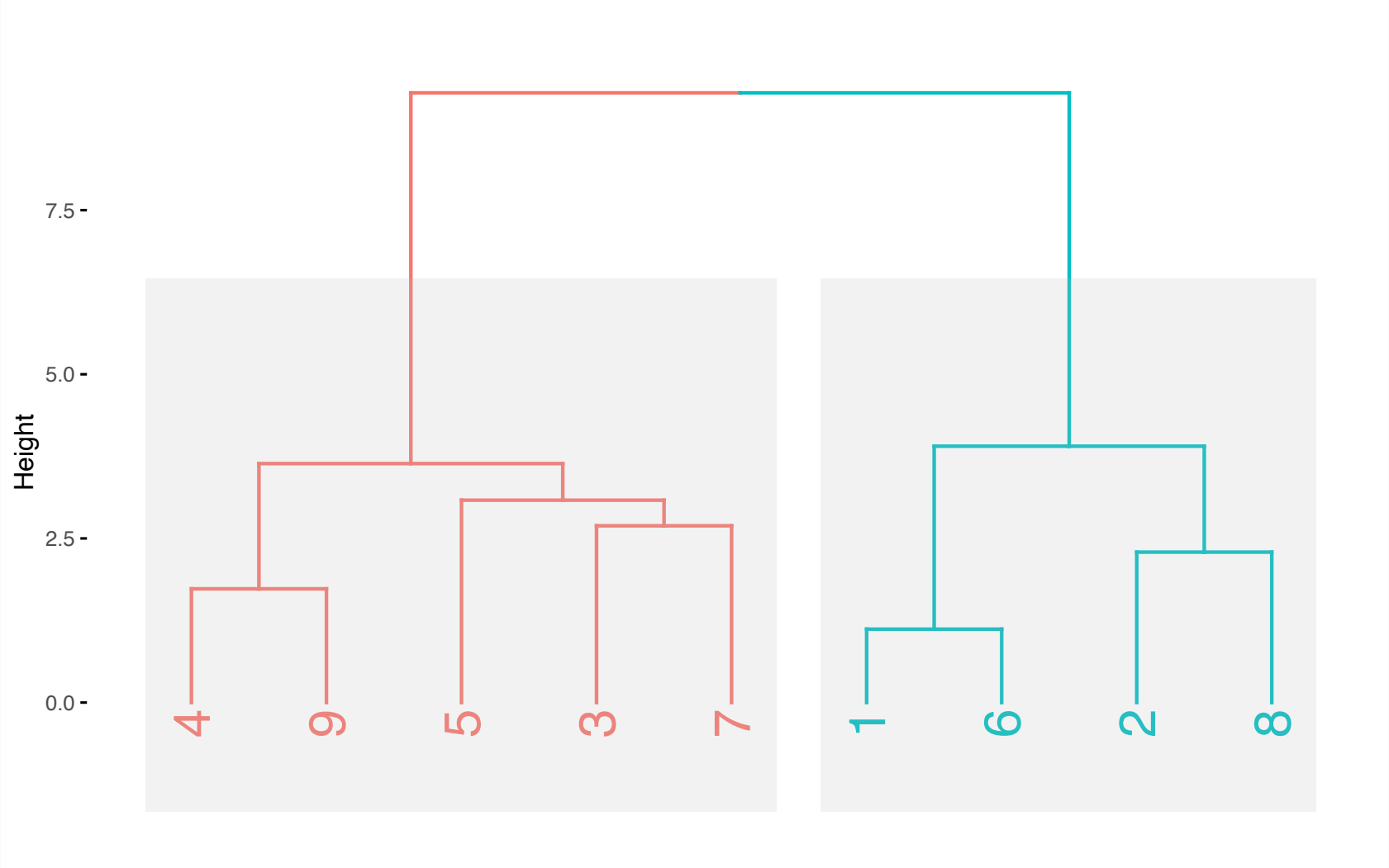
```
##      [,1] [,2]
## [1,]   -1   -6
## [2,]   -4   -9
## [3,]   -2   -8
## [4,]   -3   -7
## [5,]   -5    4
## [6,]    2    5
## [7,]    1    3
## [8,]    6    7
```

R 中的层次聚类

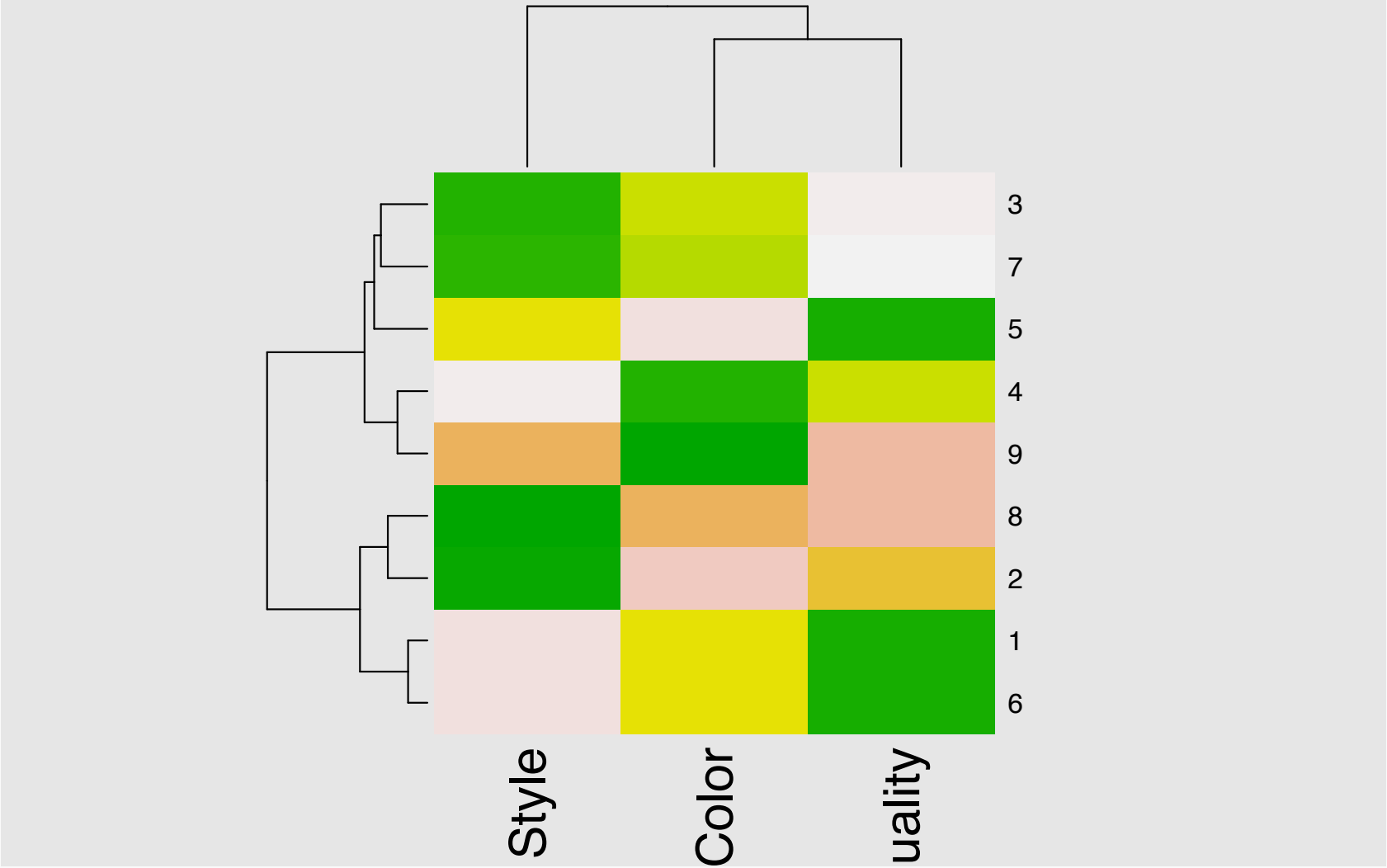
```
coat.scores.hc <- hclust(dist(coat.scores))
plot(coat.scores.hc)
```



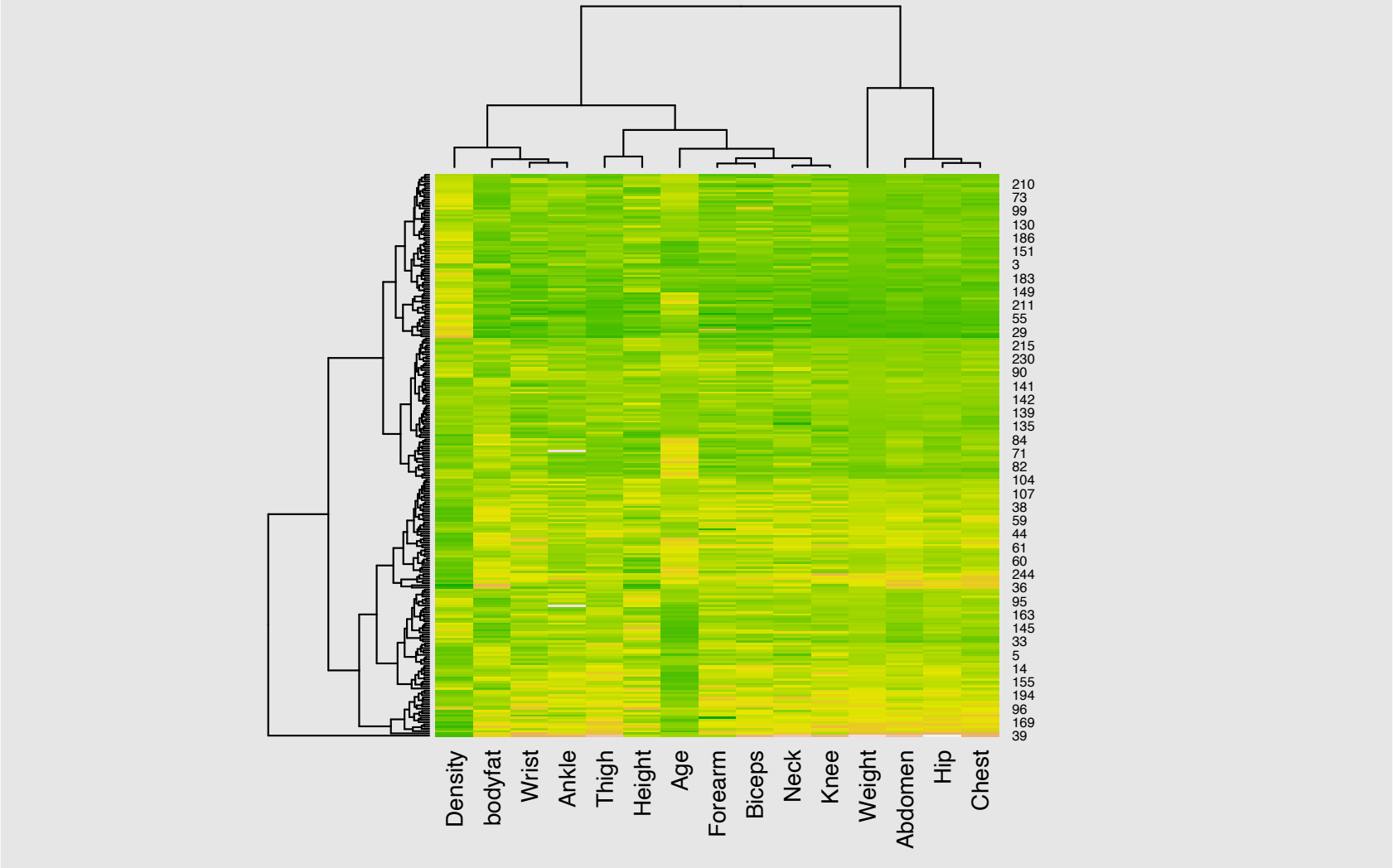
R 中的层次聚类



热力图



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

变量聚类

- 注意:至今为止我们理解了样本的聚类。
- 有时候变量聚类非常有用（用户画像)。
- 变量之间的相似性度量：相关系数。
- 变量类之间的聚合指数：最大系数法、最小系数法。

其他常用的聚类算法

机器学习中的无监督分类方法

1. 单指标分裂聚类法
2. 基于密度的聚类算法
3. 自组织图聚类（Self-Organized Maps）
4. 混合模型聚类
5. …

Outline

1 为什么聚类?

2 如何度量相似性?

3 如何聚类?

4 作业

1 下表给出六种精神治疗药物的三种临床测量指标数据，请利用分层聚类做聚类分析（分别采用最短距离法和最长距离法）。

药物	吸入量	疗效	依赖性
速可眠	5	9	20
LSD	6	11	2
安定	4	5	20
吗啡	6	9	46
仙人球毒碱	5	7	1
酒精	3	1	12

2 请采用聚类分析的方法，对污染数据.xlsx 进行聚类分析。并解释各类的意义。