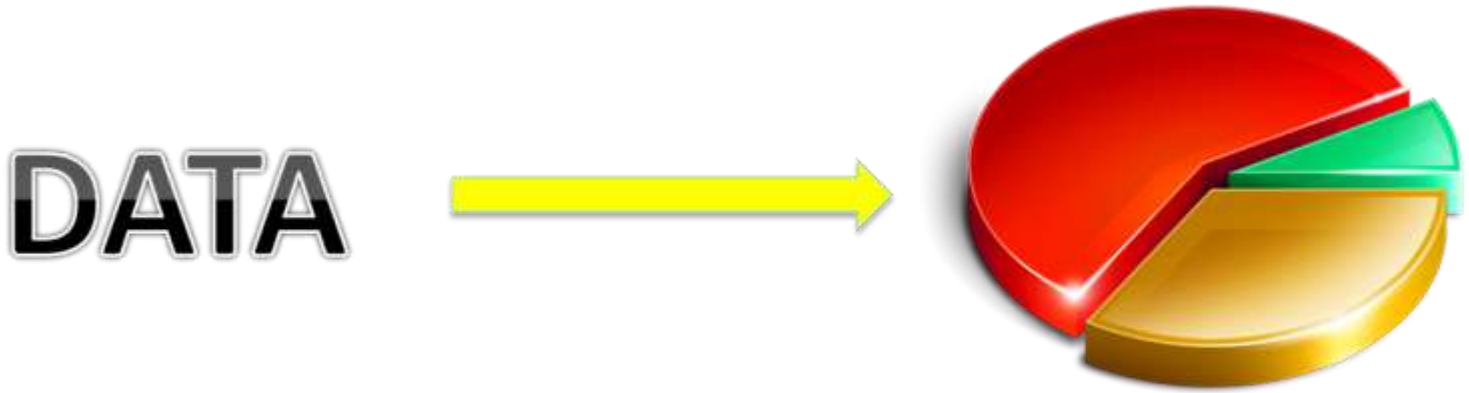
中国人民大學
RENMIN UNIVERSITY OF CHINA

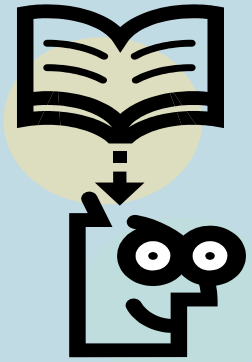
“正大杯”第12届市调大赛公益培训课

描述性统计分析与可视化



中国人民大学统计学院
中国调查与数据中心
吴翌琳

课程大纲

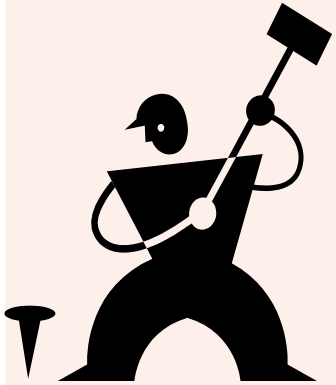


基础

- 基本概念

工具

- 可视化工具

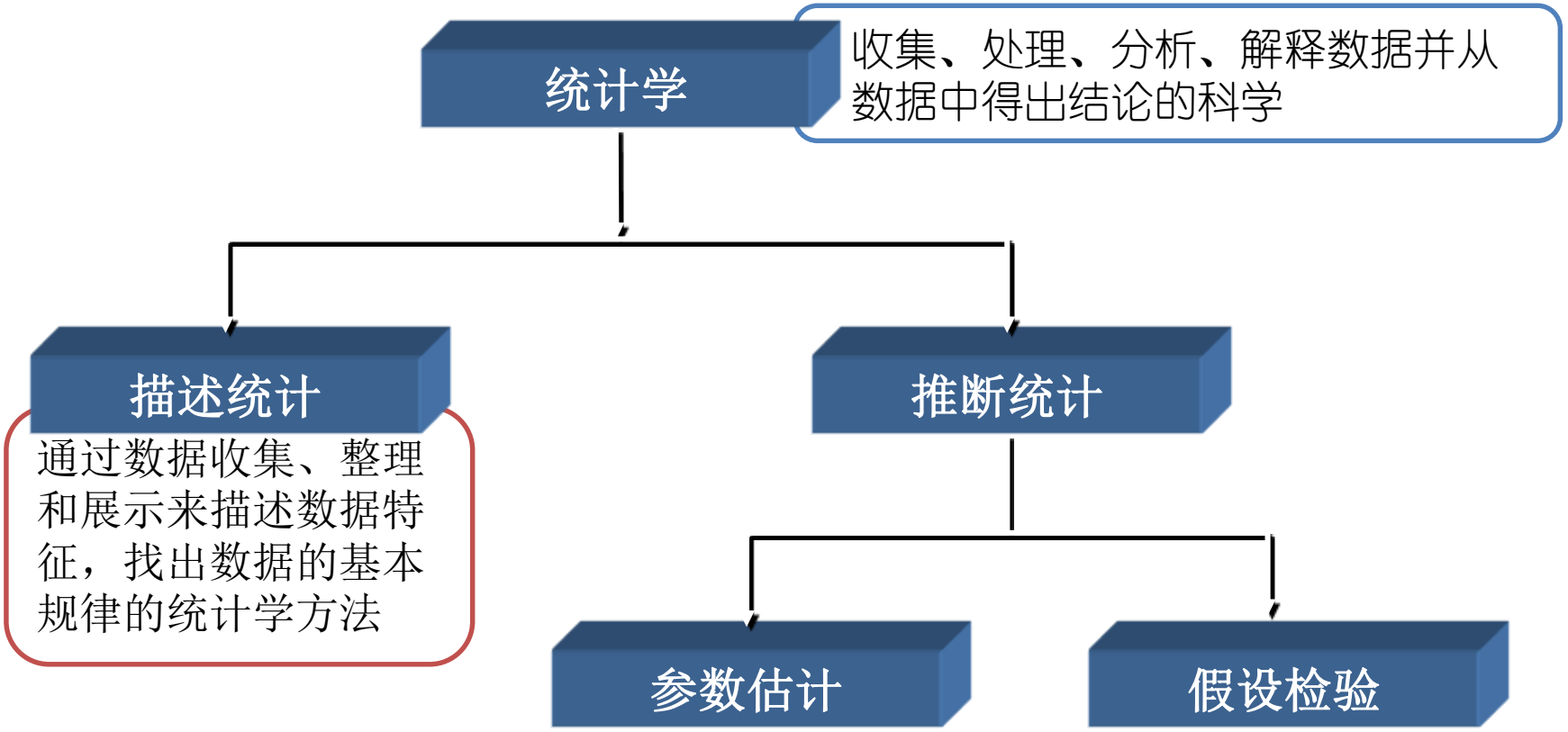


实践

- 图形设计与规范

1 基本概念

统计学方法架构


统计学
收集、处理、分析、解释数据并从数据中得出结论的科学
描述统计
通过数据收集、整理和展示来描述数据特征，找出数据的基本规律的统计学方法
推断统计
参数估计
假设检验

为什么要进行描述统计

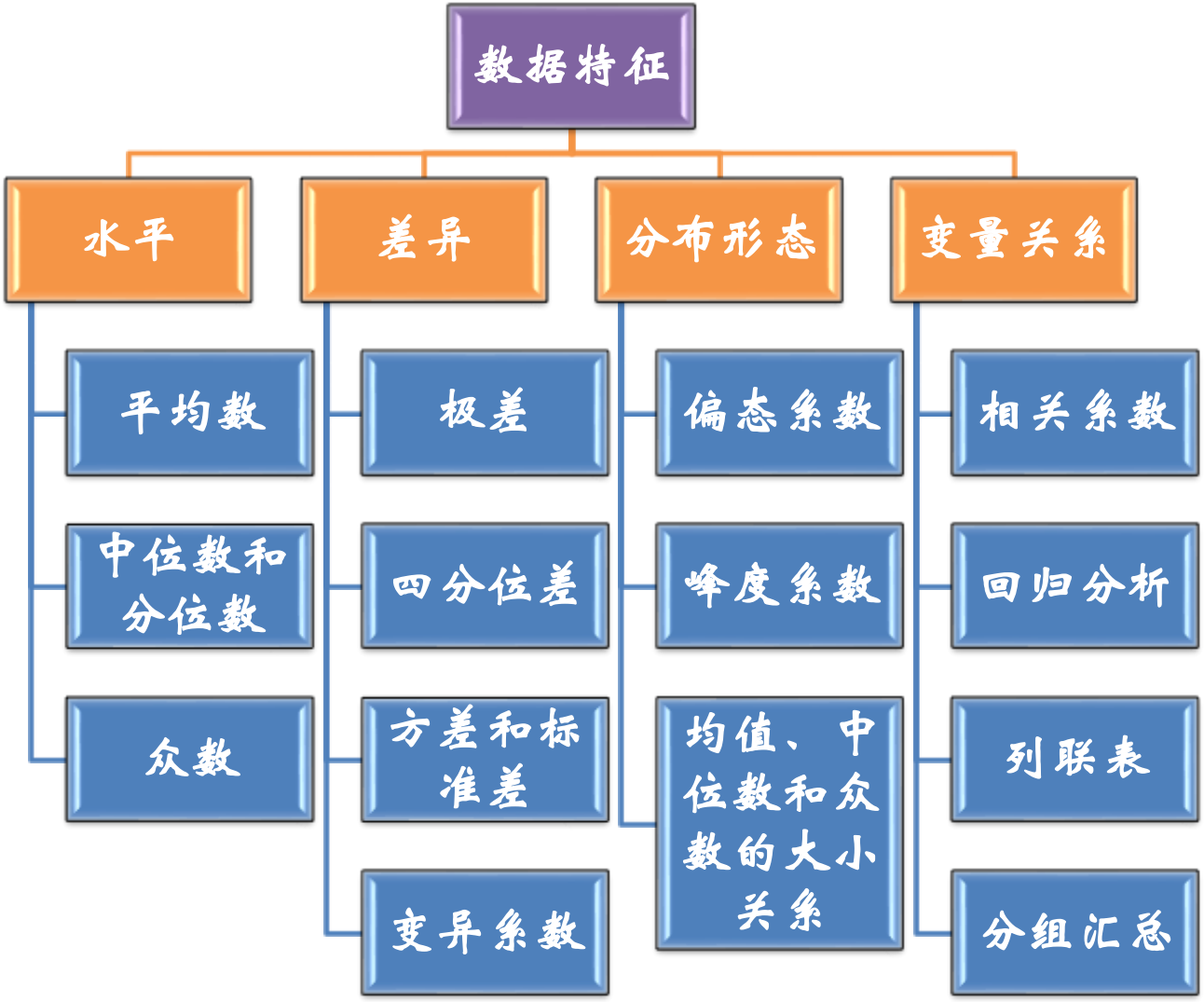
	A	B	C	D
1	性别	网购次数	网购金额	满意度
2	女	6次以上	148	满意
3	男	3-5次	4785	不满意
4	女	3-5次	2080	不满意
5	男	3-5次	3725	中立
6	女	1-2次	3085	满意
7	男	6次以上	4168	不满意
8	男	3-5次	2828	不满意
9	男	3-5次	4651	中立
10	女	3-5次	847	不满意
11	女	6次以上	4148	中立
12	男	3-5次	4270	中立
13	女	1-2次	4029	不满意
14	男	1-2次	347	不满意
15	男	3-5次	3383	中立
16	女	3-5次	3406	中立
17	女	3-5次	2615	不满意
18	男	3-5次	3606	满意
19	女	3-5次	4303	不满意
20	女	3-5次	1894	满意
21	女	6次以上	1065	满意
22	女	3-5次	4841	满意
23	男	3-5次	4202	满意
24	男	1-2次	237	不满意
25	女	6次以上	869	满意
26	女	3-5次	1094	不满意
27	女	3-5次	2050	不满意
28	女	3-5次	4946	不满意
29	男	3-5次	4880	中立
30	女	3-5次	1110	中立
31	男	1-2次	2986	中立
32	女	6次以上	1332	不满意

- 是否能直接回答以下问题：
- 1、在本次调查中，谁的网购花费最高？最低？
- 2、哪些消费者满意度更高？
- 3、大部分消费者每月的网购次数是多少？
- 4、不同性别的消费者网购情况有无差异？消费满意度有无差异？

如何进行描述统计

- 用统计量描述数据
- 用图表描述数据

用统计量描述数据


数据特征
水平
差异
分布形态
变量关系
平均数
中位数和分位数
众数
极差
四分位差
方差和标准差
变异系数
偏态系数
峰度系数
均值、中位数和众数的大小关系
相关系数
回归分析
列联表
分组汇总

用统计量进行描述统计的实现

- R中的summary函数、table函数、pastecs包的stat.desc函数等

```
##  性别          网购次数      网购金额            满意度        index
##  男: 840          :  4   Min.   :      11          :  3   Min.   :   1.0
##  女:1160   1-2次  :615   1st Qu.:    1262   不满意:799   1st Qu.: 500.8
##            3-5次  :804   Median :    2491   满意  :518   Median :1000.5
##            6次以上:577   Mean   :    7518   中立  :680   Mean   :1000.5
##                          3rd Qu.:    3780                3rd Qu.:1500.2
##                          Max.   :10000000                Max.   :2000.0
##                          NA's   :5
```

- Excel中的“汇总统计
- SPSS中的“Analyze -> Descriptive Statistics -> Descriptive”
- Python中的NumPy和SciPy
- ……

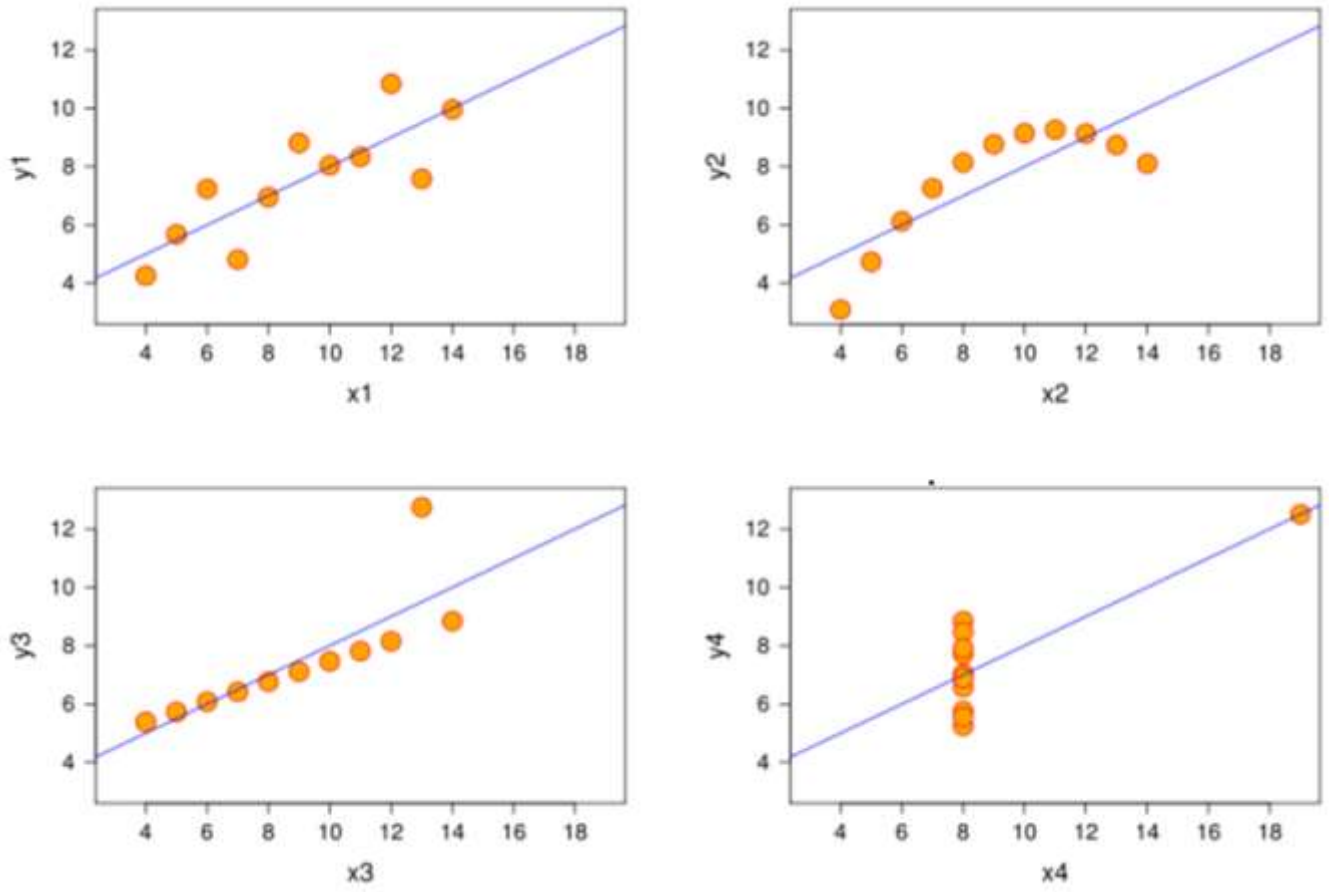
找出数据特征

Anscombe's Quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

- 这四组数据中：
- x值的平均数都是9.0，y值的平均数都是7.5
- x值的方差都是10.0，y值的方差都是3.75
- 它们的相关度都是0.816
- 线性回归线都是y=3+0.5x

而事实上...



- 1973年，统计学家F.J. Anscombe构造出了四组奇特的数据。它告诉人们，在分析数据之前，描绘数据所对应的图像有多么的重要。

数据可视化

- 借助于图形化手段，清晰有效地传达与沟通信息。（维基百科）
- 数据可视化起源于1960年计算机图形学，那时候人们使用计算机创建图形图表，可视化提取出来的数据，可以将数据的各种属性和变量呈现出来。
- 随着计算机硬件的发展，人们创建更复杂规模更大的数字模型，于是乎发展了数据采集设备和数据保存设备，而此时也需要更高级的计算机图形学技术及方法来创建这些规模庞大的数据集。
- 随着数据可视化平台的拓展，应用领域的增加，表现形式的不断变化，以及增加了诸如实时动态效果、用户交互使用等，数据可视化像所有新兴概念一样边界不断扩大。
- 可视化的美丽之一在于简单，表现清晰

统计制图三大要素

- **信息**
  - 海量 复杂 高维 清理 统计
- **设计**
  - 视觉 交互 简介 适度
- **沟通**
  - 直观 高效 传递信息 发现知识

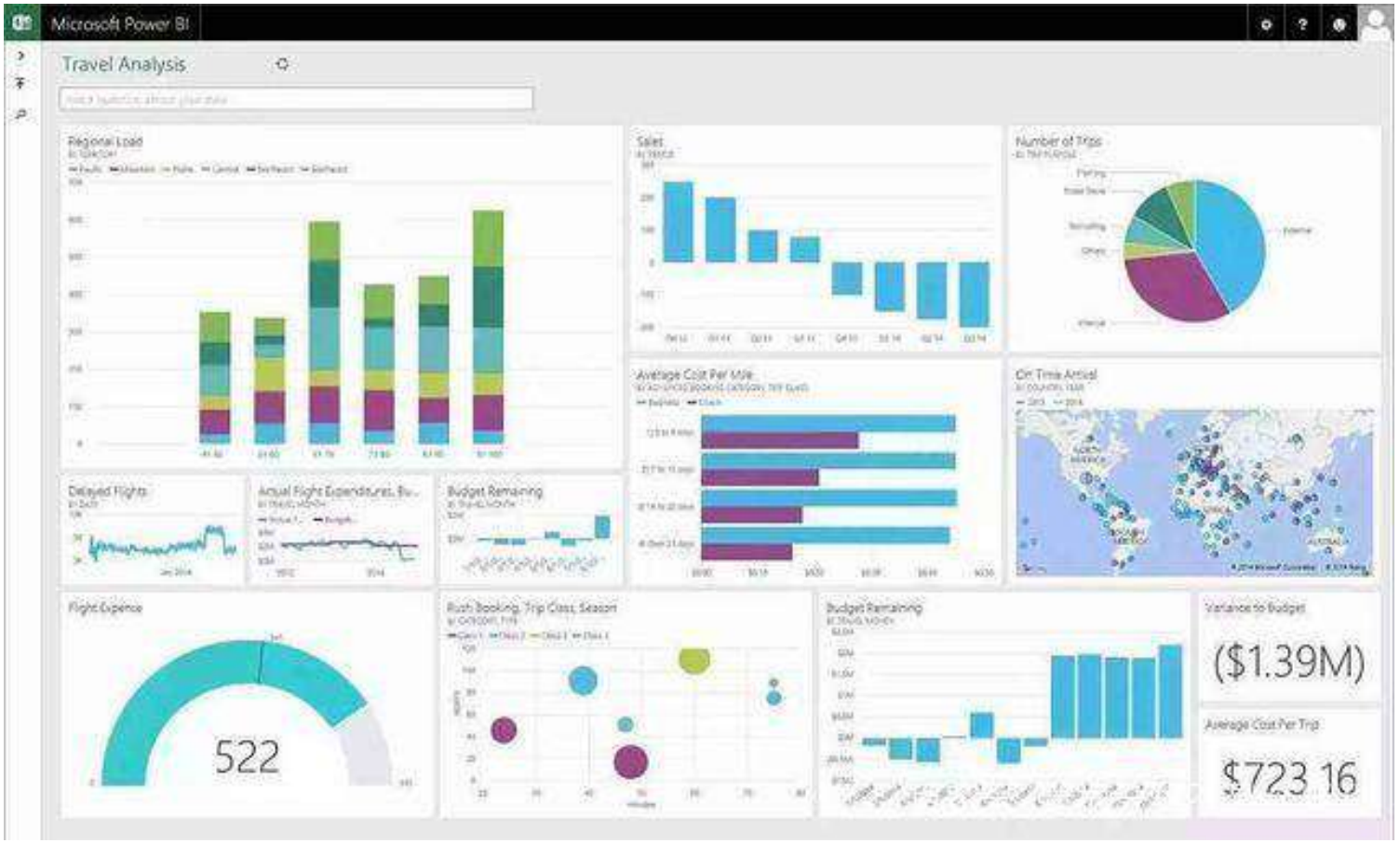
2 可视化的工具

可视化工具

- Excel及其插件
- Echart
- R
- Python
- Tableau
- 数据可视化平台：网易有数、百度图说.....

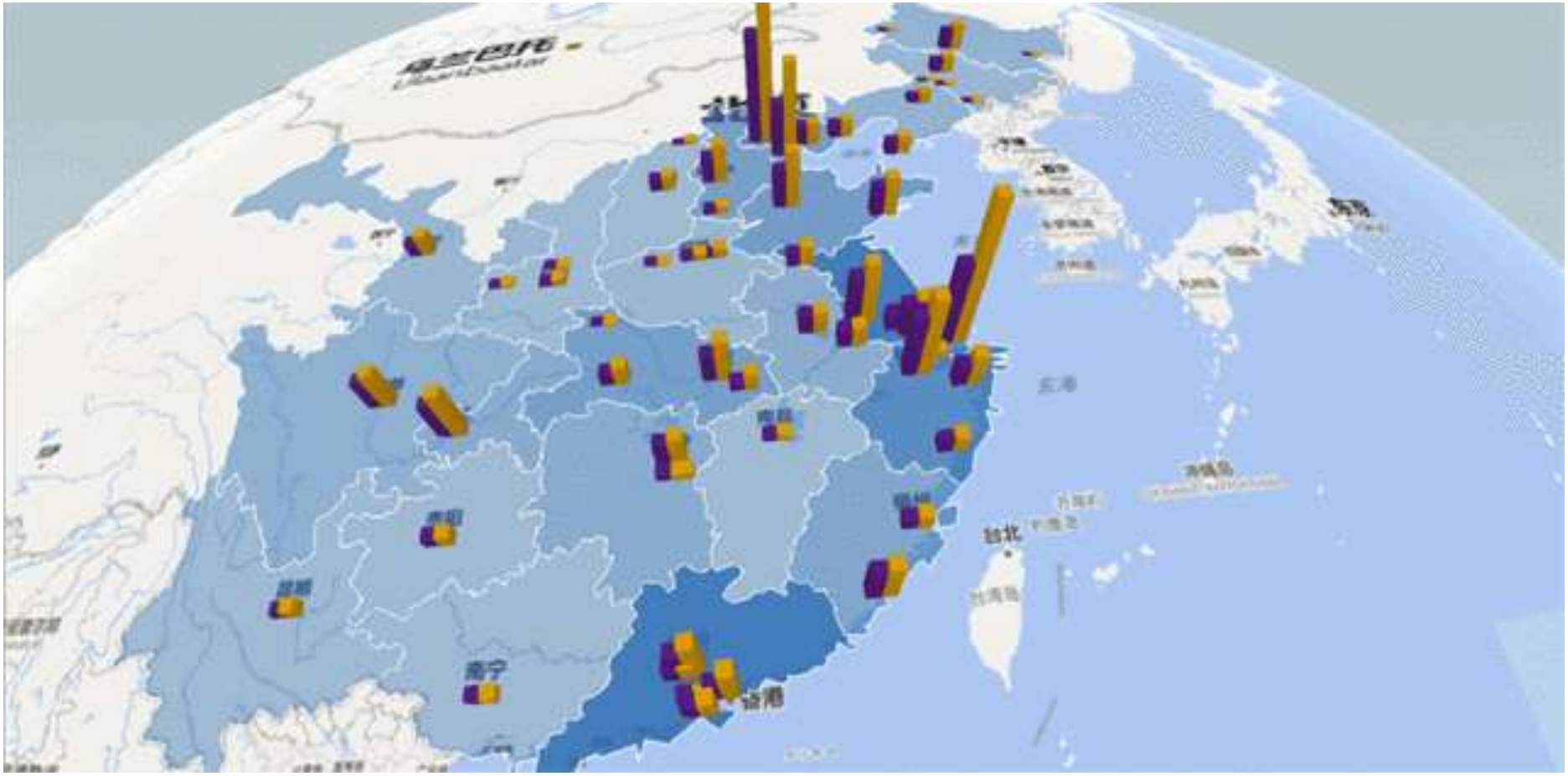
PowerBI

- 软继Excel之后推出的BI产品，可以和Excel无缝连接使用，创建个性化的数据看板。



Power Map

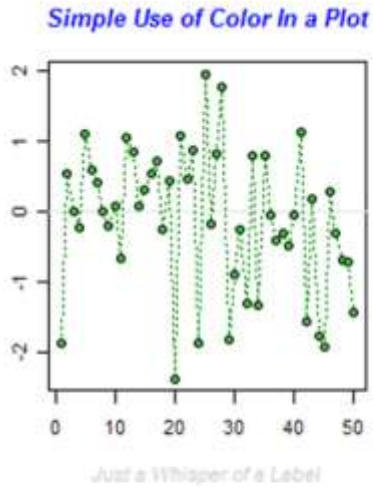
- Excel中推出的一个功能强大的加载项，结合Bing地图，支持用户绘制可视化的地理和时态数据，并用3D方式进行分析。同时，用户还可以使用它创建视



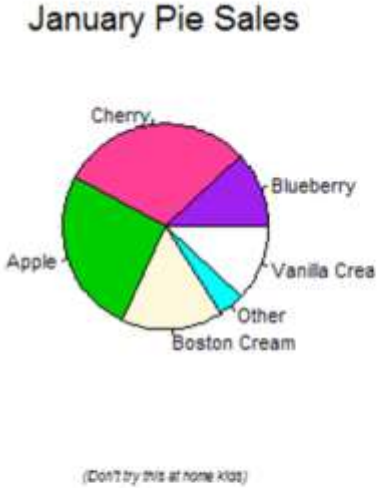
Echart

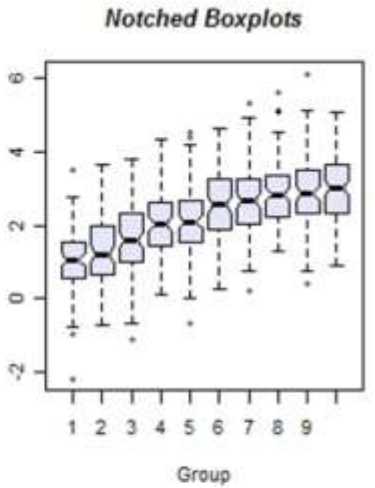
- 一个纯Javascript的数据可视化库，百度的产品，常应用于软件产品开发或网页的统计图表模块。可在

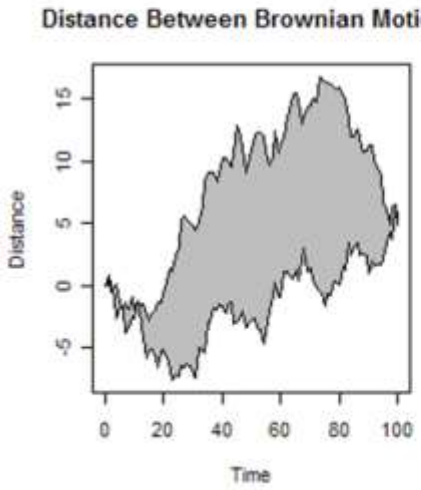
R-graphic


Simple Use of Color In a Plot
Just a Whisper of a Label

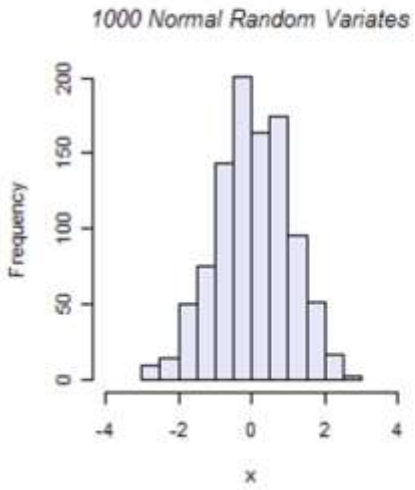

A Sample Color Wheel
(Use this as a test of monitor linearity)

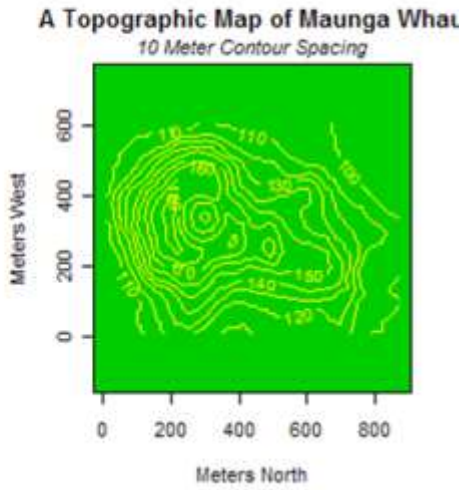

January Pie Sales
Cherry
Blueberry
Apple
Vanilla Crea
Other
Boston Cream
(Don't try this at home kids)


Notched Boxplots
Group

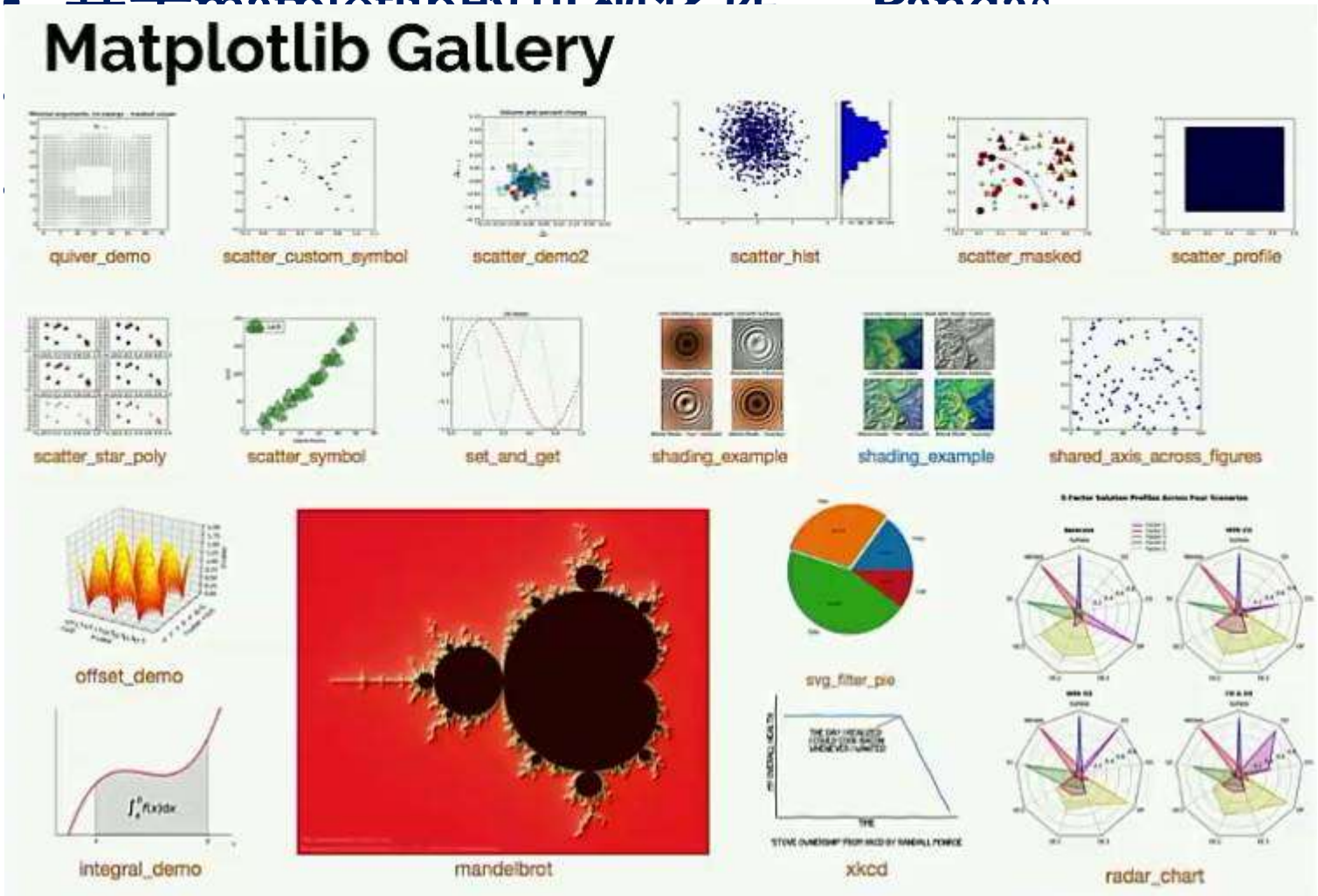

Distance Between Brownian Motions
Distance
Time


The Level of Interest in R
Jan
Mar
May
Jul
Sep
Nov
1996


1000 Normal Random Variates
Frequency
x

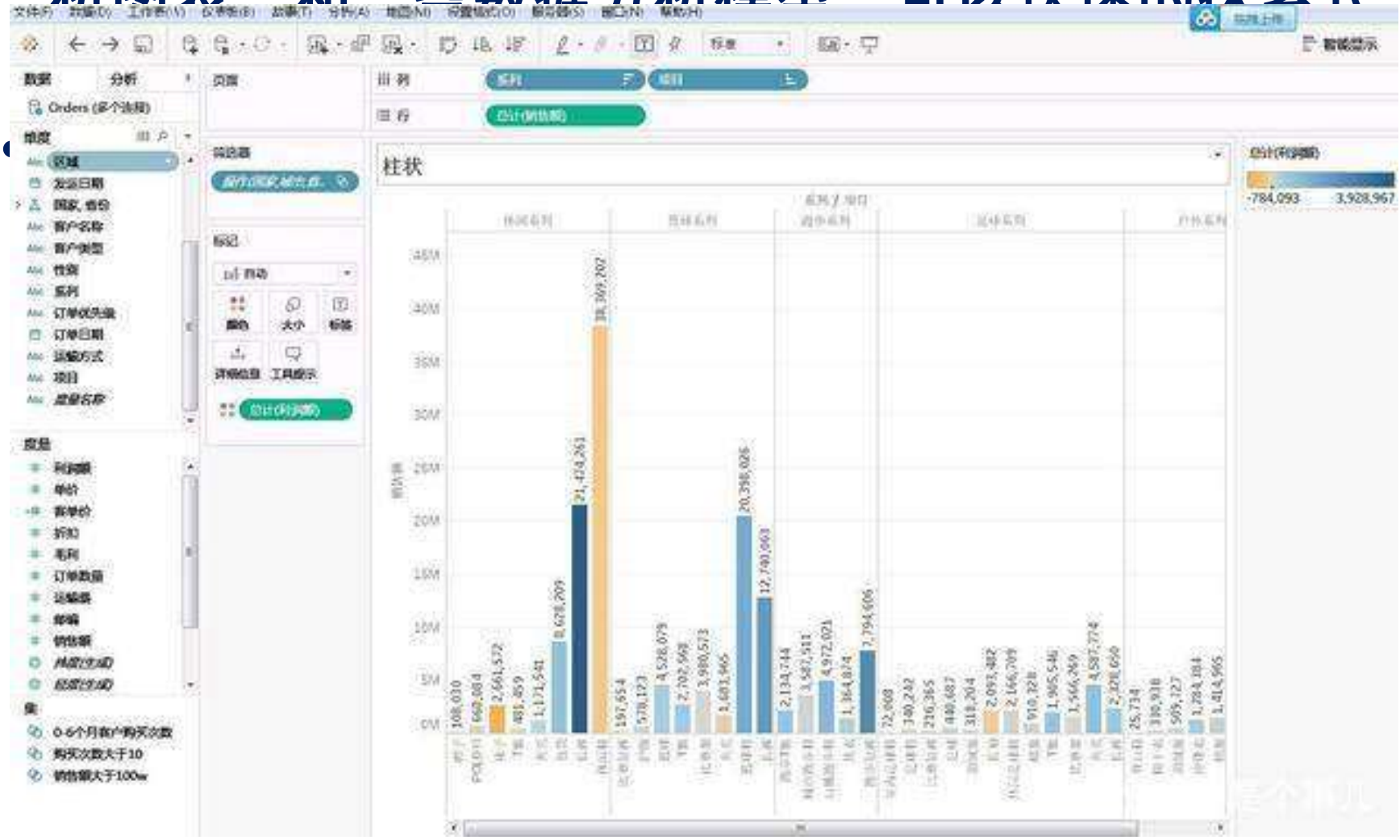

A Topographic Map of Maunga Whau
10 Meter Contour Spacing
Meters West
Meters North

- python可视化库可以大致分为几类：
  - 基于matplotlib的可视化库 Pandas



Tableau

- 几乎是数据分析师人人会提的工具，内置常用的分析图表，和一些数据分析模型，可以快速的探索式



3 统计制图的流程与设计

统计制图的七个阶段

- 获取　　数据来源
- 分析　　数据结构
- 过滤　　关注信息
- 挖掘　　统计辨析
- 表述　　视觉模型
- 修饰　　清晰易读
- 交互　　操作控制

分析

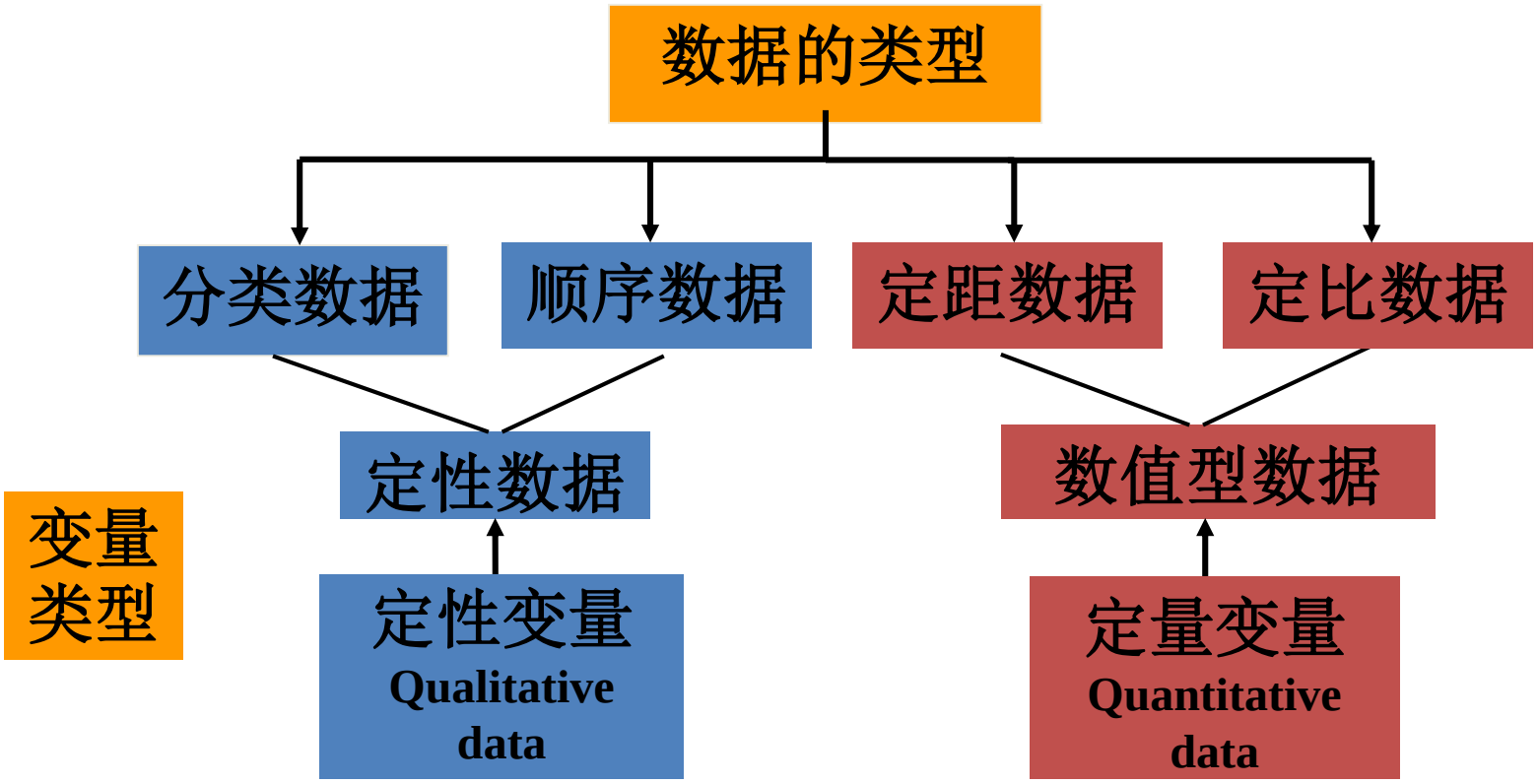
- 想表达什么？
- 想解决什么样的问题？
- 是否可以实现？
- 谁是这个数据的使用者？
- 他们需要什么样的数据？

如何表现数据之间的关系?

- 通过不同的位置表示关系
- 尝试使用多种坐标轴
- 考虑如何定义其格式
- 使用不同类型的颜色
- 使用适当的表现属性呈现数值

3.1 图表类型的选择

变量与数据的分类



- 不同尺度和变量所含信息量不同
- 不同尺度/变量对应不同的数据显示方法和分析方法

举例

性别： 男 ← 分类数据

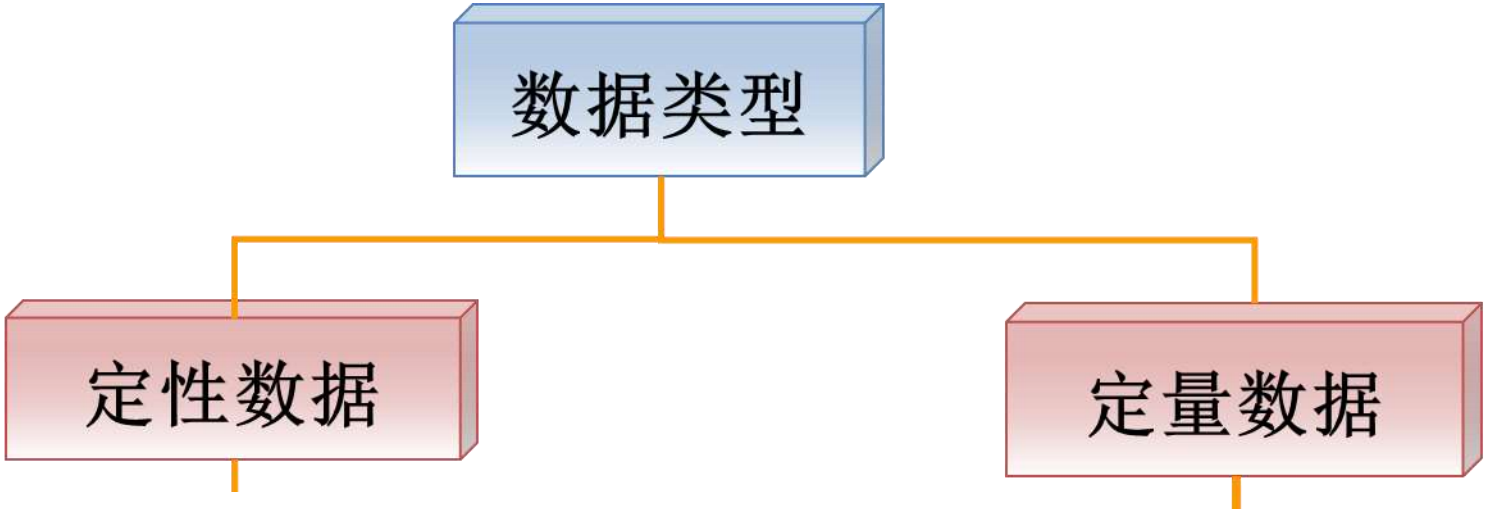
健康状况： 良好 ← 顺序数据

出生年份：1980 ← 定距数据

体重： 134.5公斤 ← 定比数据

精确程度

根据数据的类型选择



3.2 定性变量数据的展示

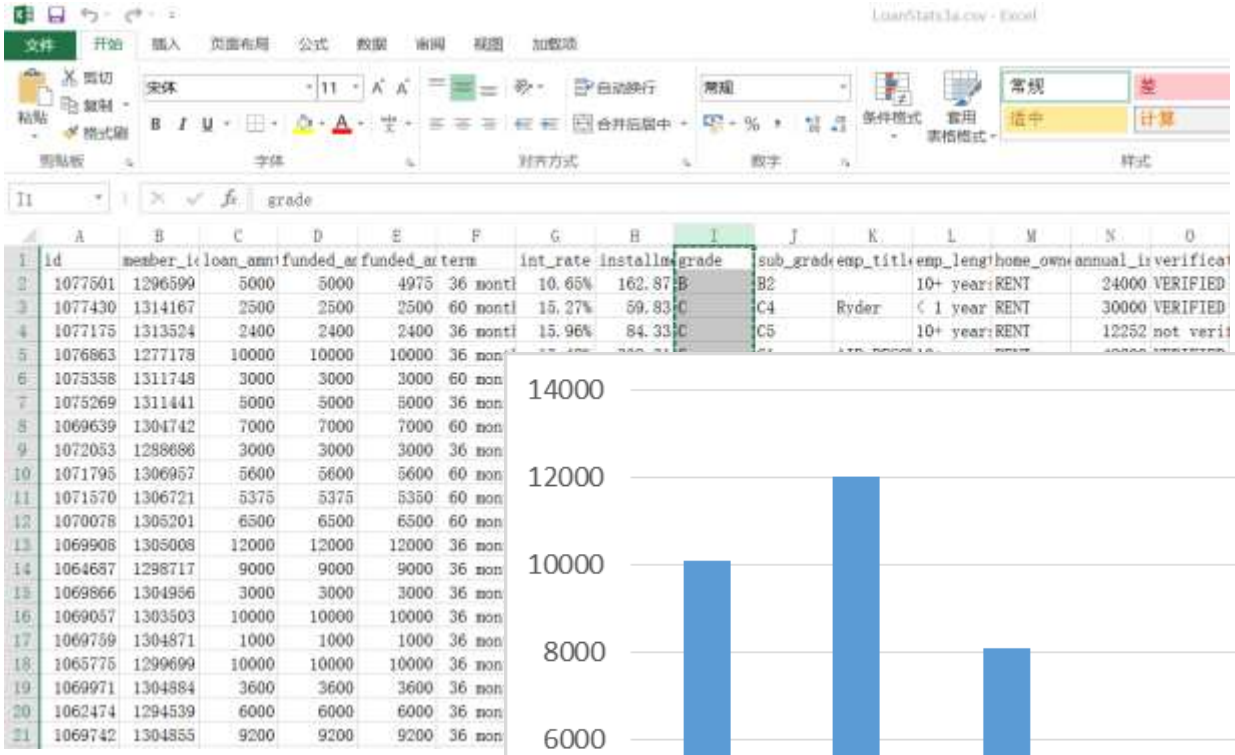
定性变量数据的展示

- 柱状图
- 点图（哑铃图）
- 饼图（环形图）
- 风玫瑰图
- 树状图
- 马赛克图
- 独立性检验P值图
- 气球图
- 热图
- 词云图
- ……

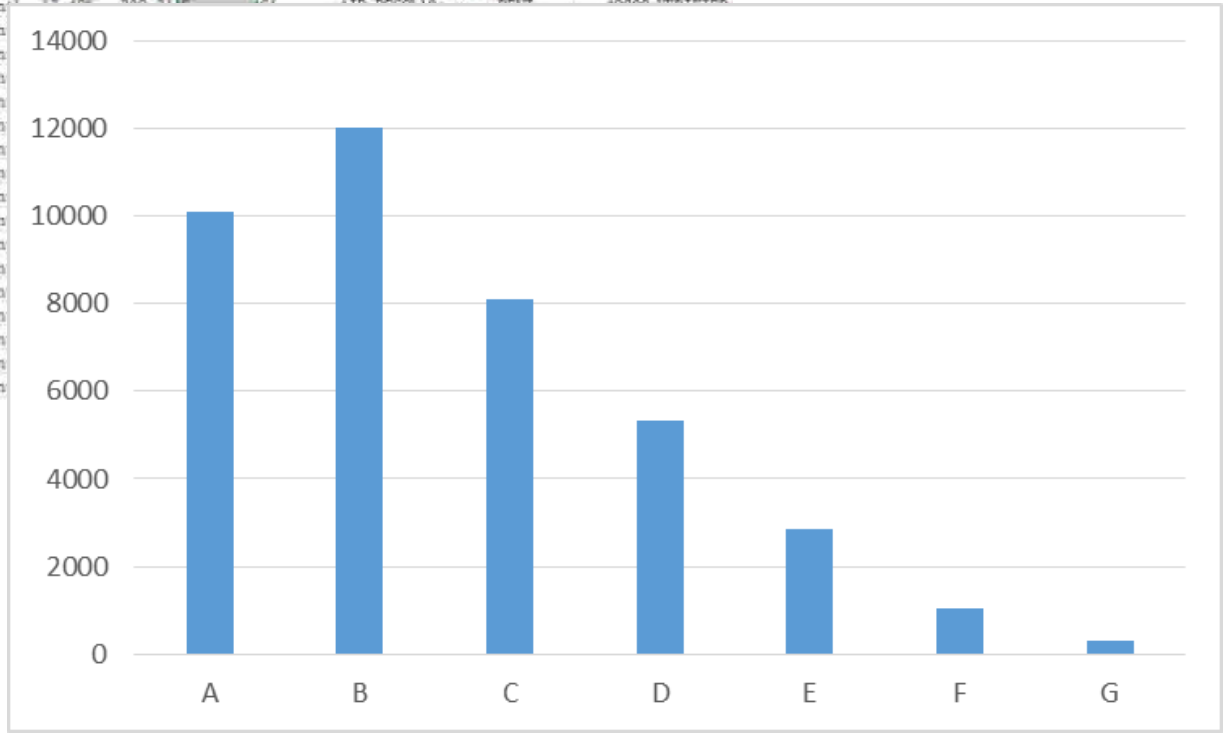
柱状图/条形图(Bar Chart)

- 用宽度相同的柱子高度或长短来表示数据变动的图形。
- 柱子的排列可以横排，也可以纵排
- 柱状图有单式、复式等形式
- 适用数据：定性数据
- 主要功能：进行类别频数、比例的展示和比较

EXCEL绘制柱状图



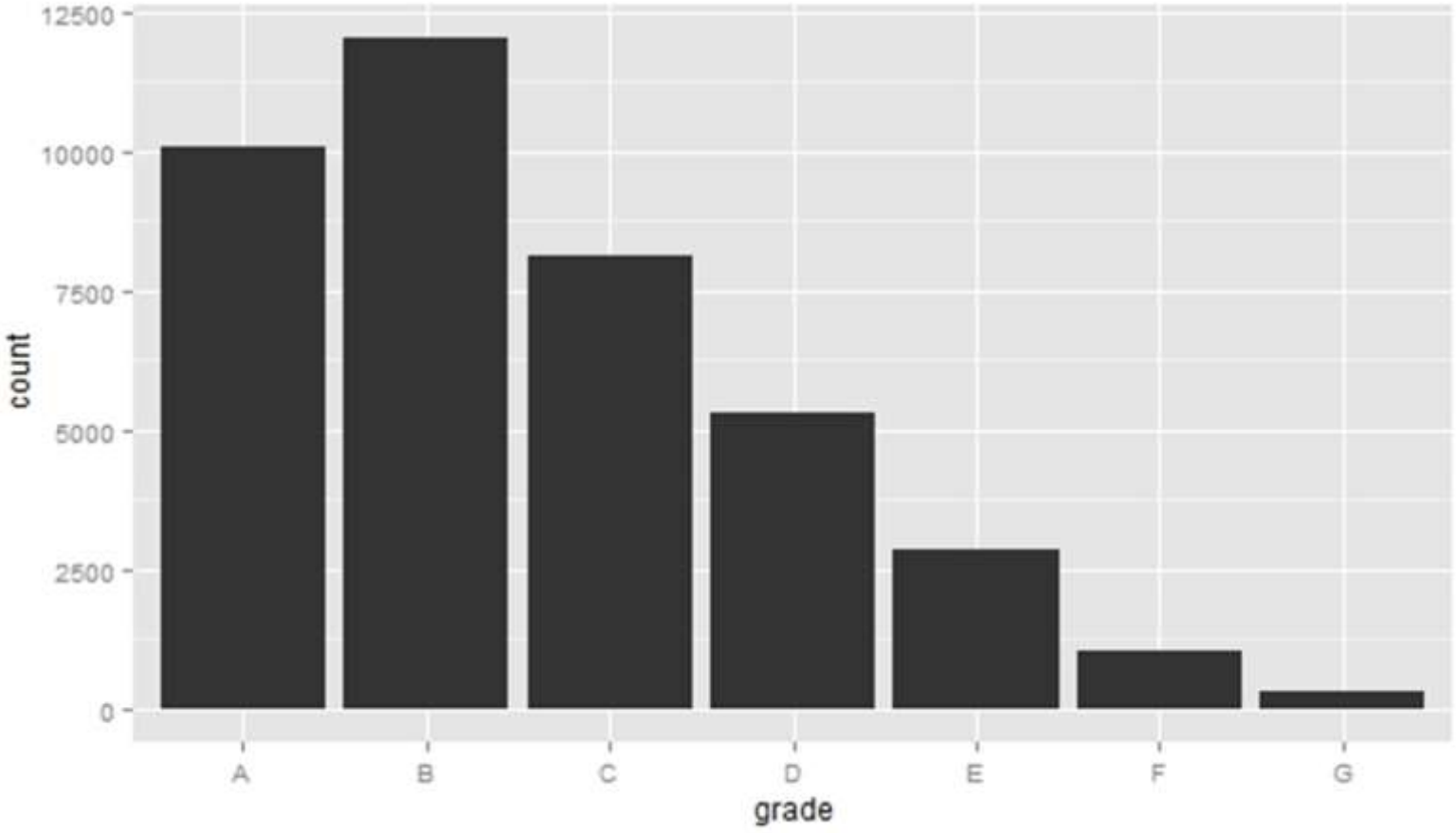
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	id	member_id	loan_amn	funded_a	funded_a	term	int_rate	installm	grade	sub_grad	emp_titl	emp_leng	home_own	annual_i	verifica
2	1077501	1296599	5000	5000	4975	36 monti	10.65%	162.87	B	B2		10+ year	RENT	24000	VERIFIED
3	1077430	1314167	2500	2500	2500	60 monti	15.27%	59.83	C	C4	Ryder	< 1 year	RENT	30000	VERIFIED
4	1077175	1313524	2400	2400	2400	36 monti	15.96%	84.33	C	C5		10+ year	RENT	12252	not veri
5	1076863	1277178	10000	10000	10000	36 mon									
6	1075358	1311748	3000	3000	3000	60 mon									
7	1075269	1311441	5000	5000	5000	36 mon									
8	1069639	1304742	7000	7000	7000	60 mon									
9	1072053	1288686	3000	3000	3000	36 mon									
10	1071795	1306957	5600	5600	5600	60 mon									
11	1071570	1306721	5375	5375	5350	60 mon									
12	1070078	1305201	6500	6500	6500	60 mon									
13	1069908	1305008	12000	12000	12000	36 mon									
14	1064687	1298717	9000	9000	9000	36 mon									
15	1069866	1304956	3000	3000	3000	36 mon									
16	1069057	1303503	10000	10000	10000	36 mon									
17	1069759	1304871	1000	1000	1000	36 mon									
18	1065775	1299699	10000	10000	10000	36 mon									
19	1069971	1304884	3600	3600	3600	36 mon									
20	1062474	1294539	6000	6000	6000	36 mon									
21	1069742	1304855	9200	9200	9200	36 mon									



R-频数柱状图

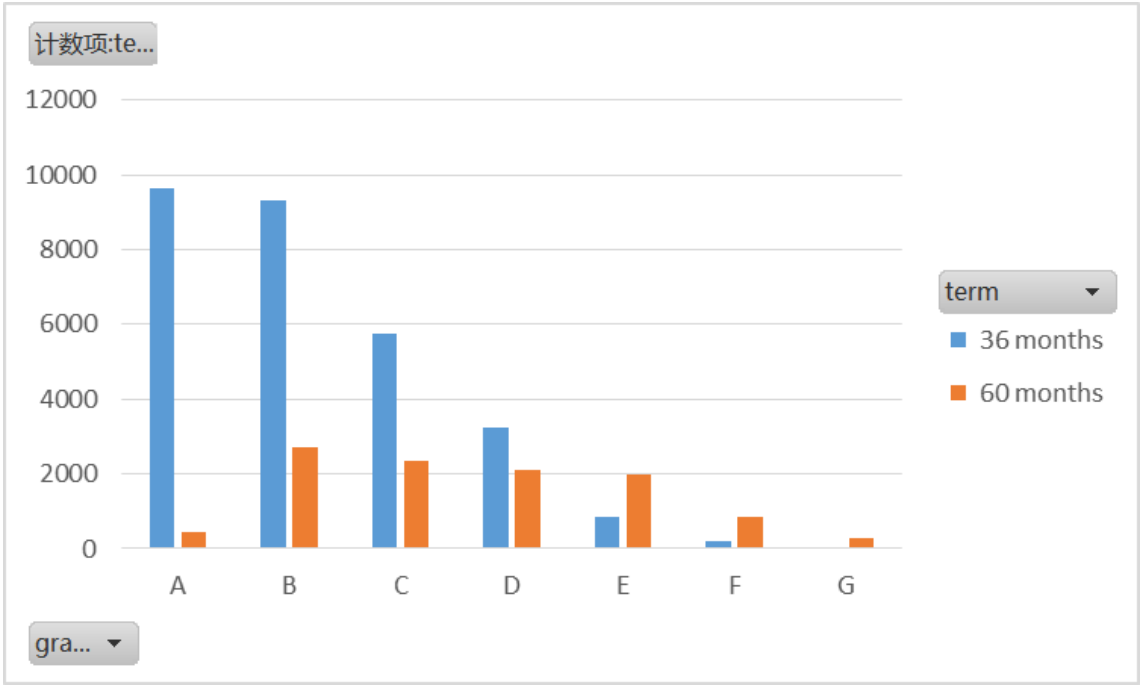
- 默认设置为stat="bin"，即自动计算每组频数

```
ggplot(data,aes(x=grade))+geom_bar()
```



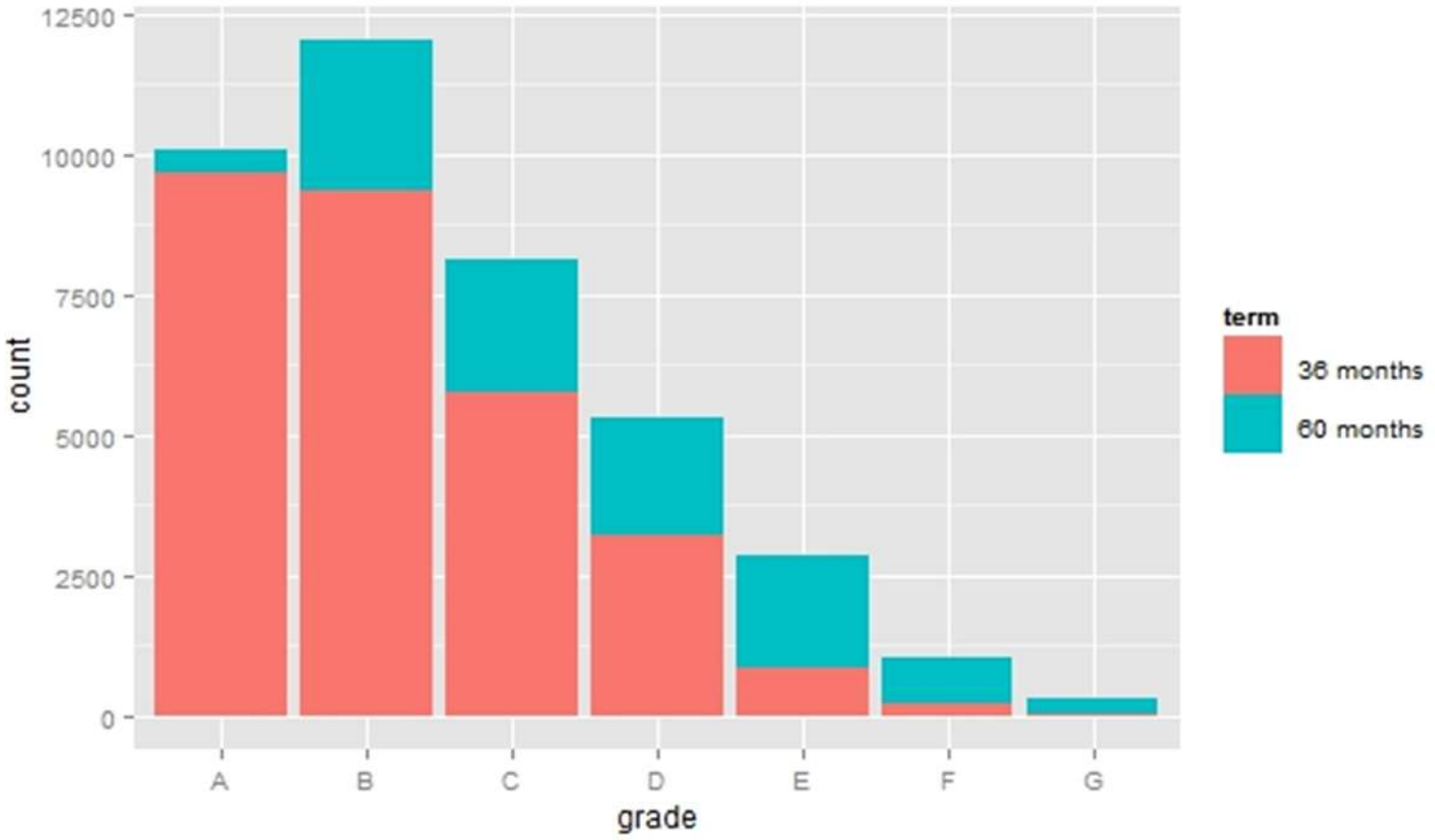
复杂柱状图

- 簇状柱状图和堆积柱状图包含两个定性变量（分类变量）
- 基本原理是用fill函数将不同类别填充为不同颜色
- 以频数柱状图为例，选取原数据中term与grade两个分类变量。



堆积柱状图

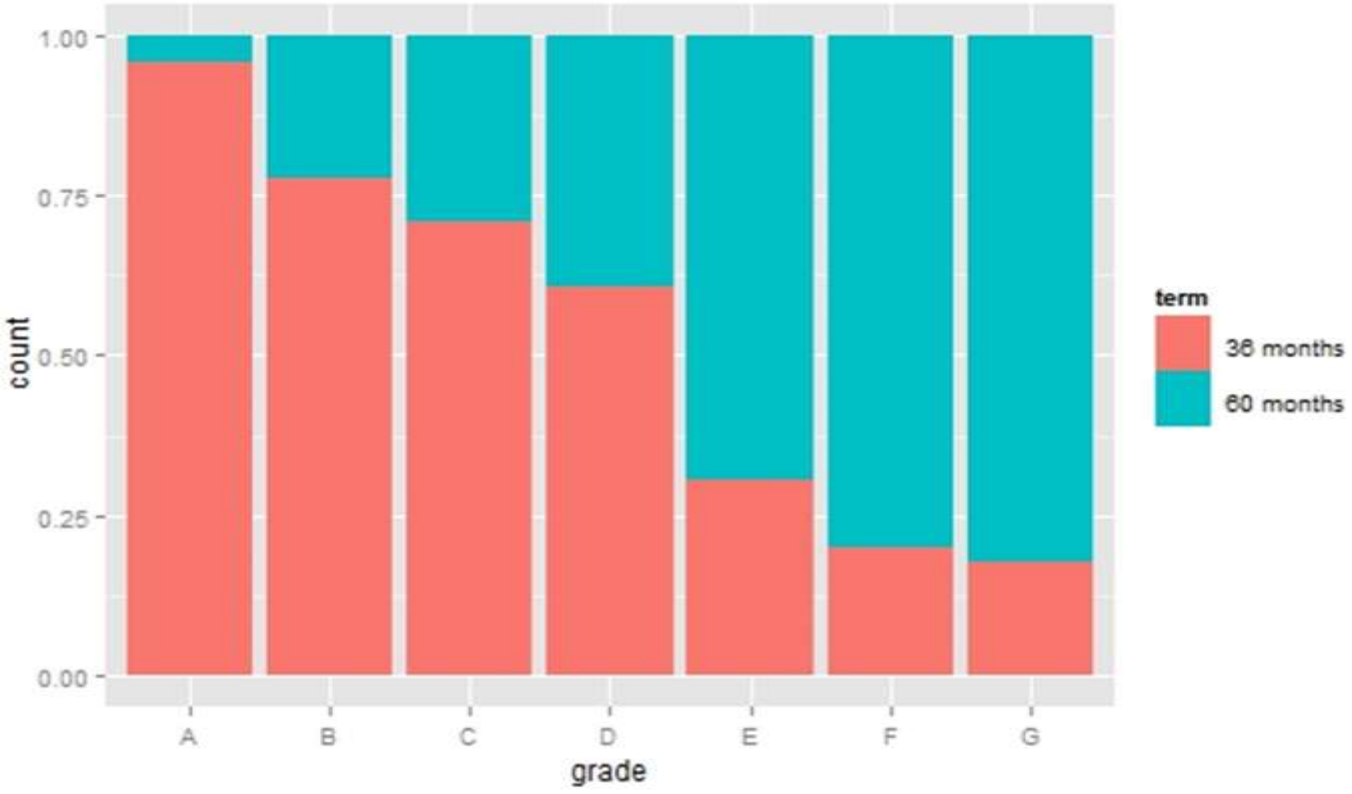
```
ggplot(data,aes(x=grade,fill=term))+geom_bar()
```



堆积柱状图的标准化

- 堆积柱状图可以将堆叠的高度标准化：position="fill"

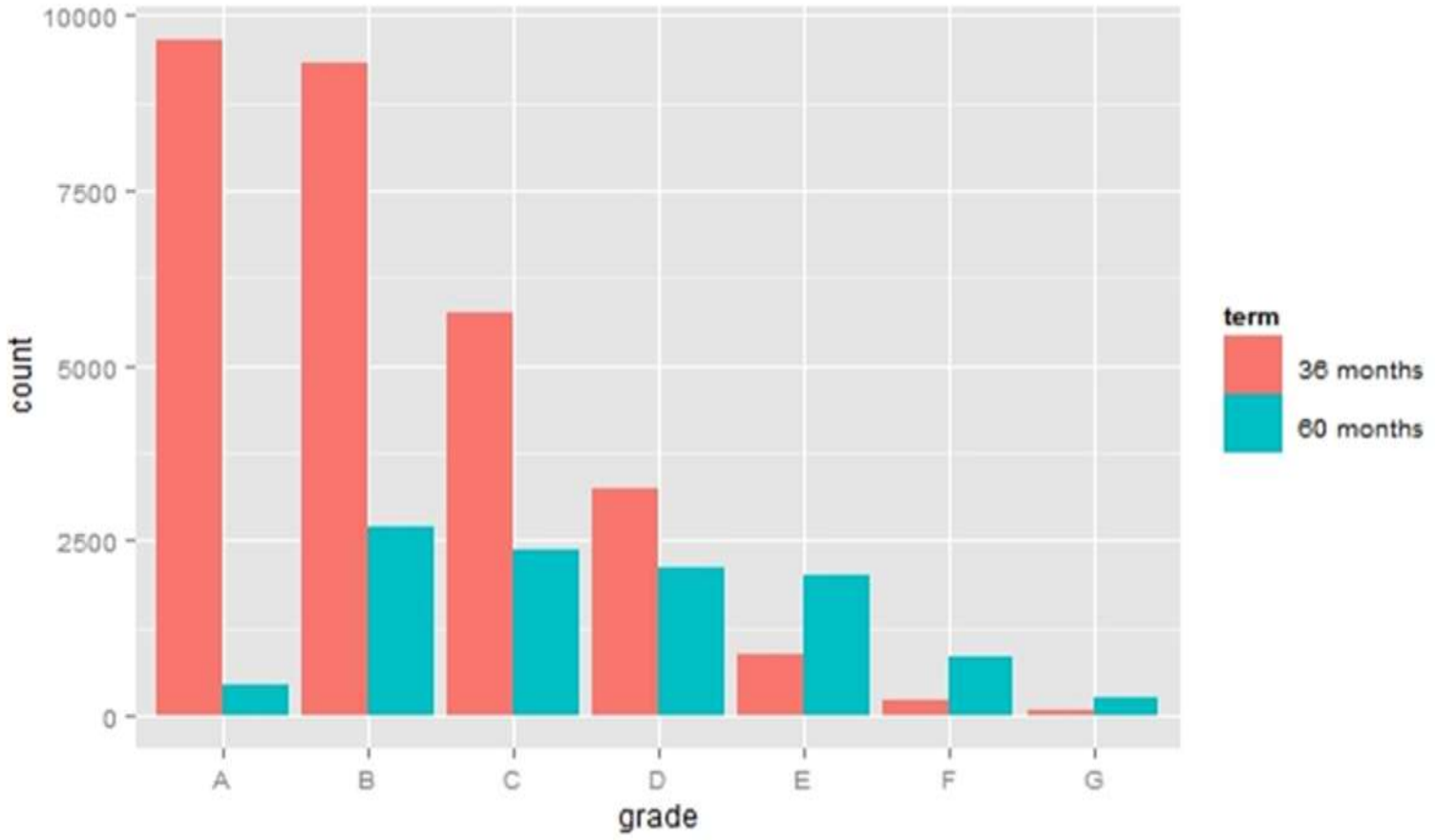
```
ggplot(data,aes(x=grade,fill=term))+geom_bar(position="fill")
```



簇状柱状图

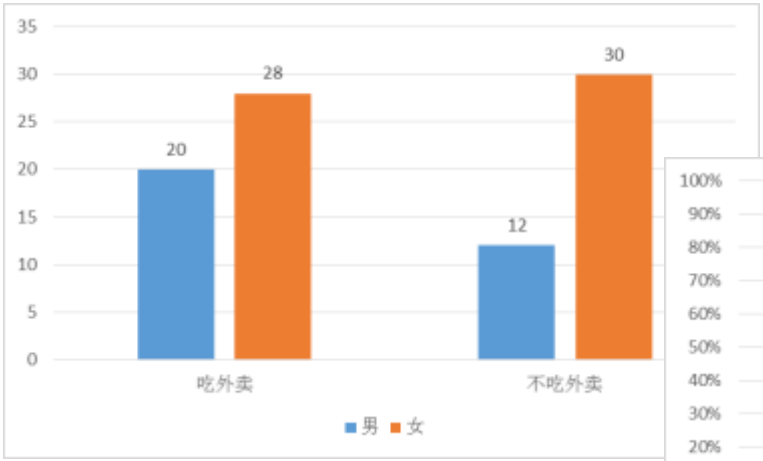
```
ggplot(data,aes(x=grade,fill=term))
+geom_bar(position="dodge")
```

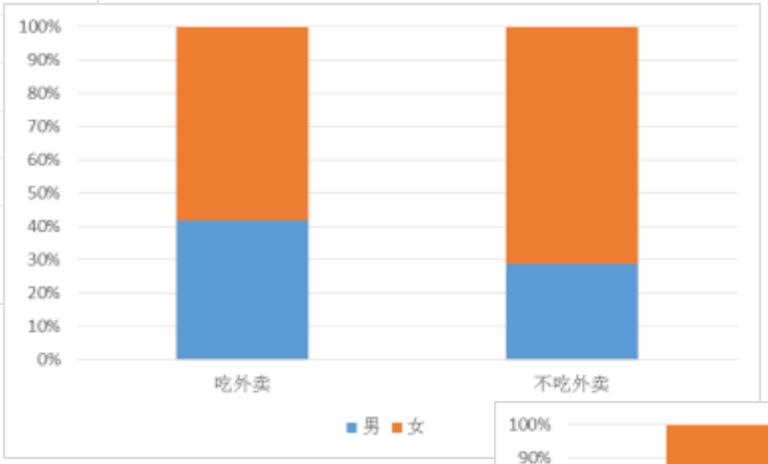
将堆积改为并列

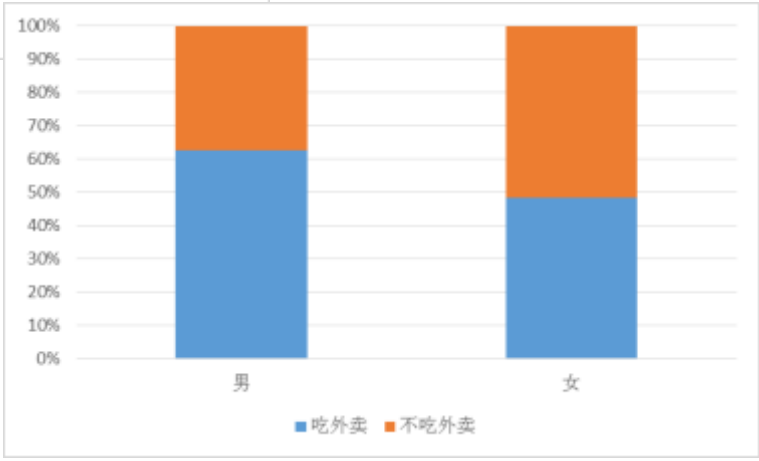


堆积or百分比堆积?

- **到底谁比较爱吃外卖?**

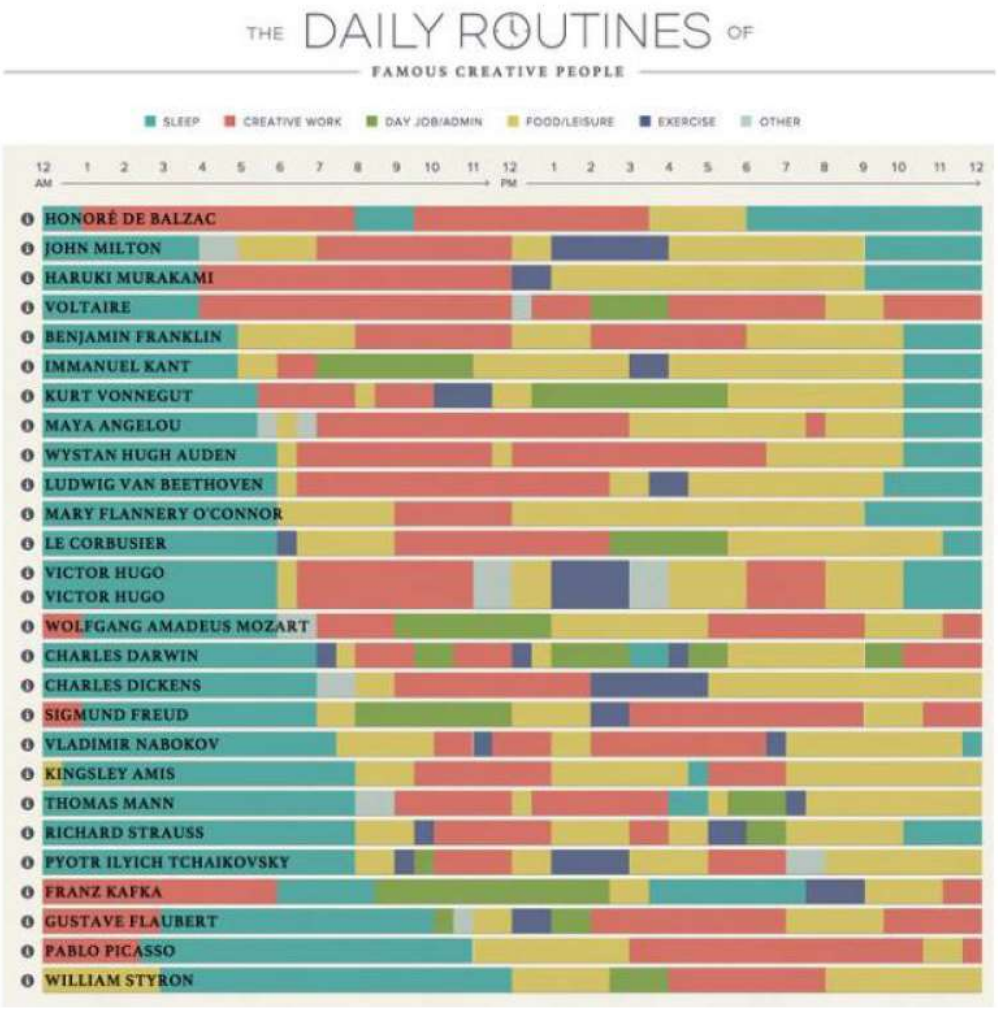






	吃外卖	不吃外卖
男	20	12
女	28	30

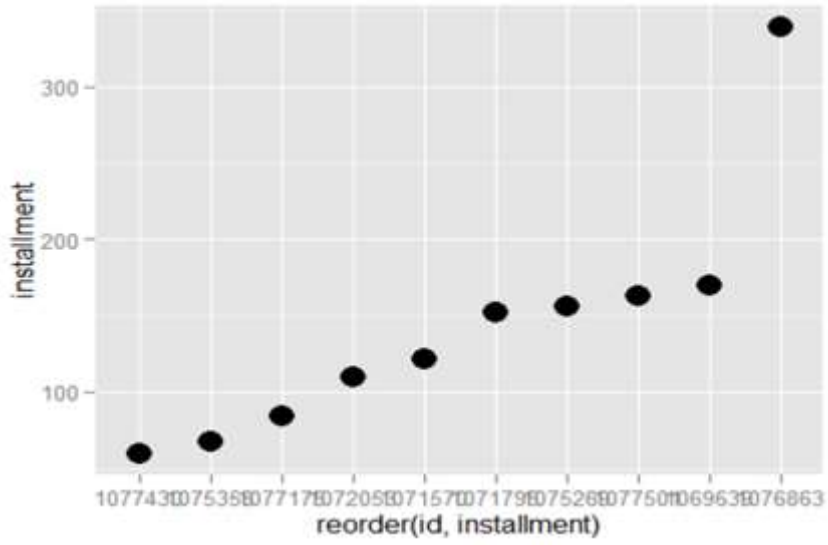
著名创意人士的时间安排



- 这个数据可视化图展示了某些著名创意人士的日程安排。横轴表示一天的时间，每一行代表一个人在一天各个时间段的活动安排
- 通过图表解读其时间和活动安排，可以发现这些著名创意人士的一些共同特点和个性区别，比如他们大部分时间都花在睡觉、创作以及进食，但是每个人创作的时间点又各有不同，有的人凌晨的创作力最强，有的人则习惯下午进行创作等。
- 这个图可以看作是横着摆放的堆积条形图。包含两种数据类型，一种是定性变量，一种是定量变量，定性变量确定该段时间是用于做什么事情，定量变量衡量时间的长短。

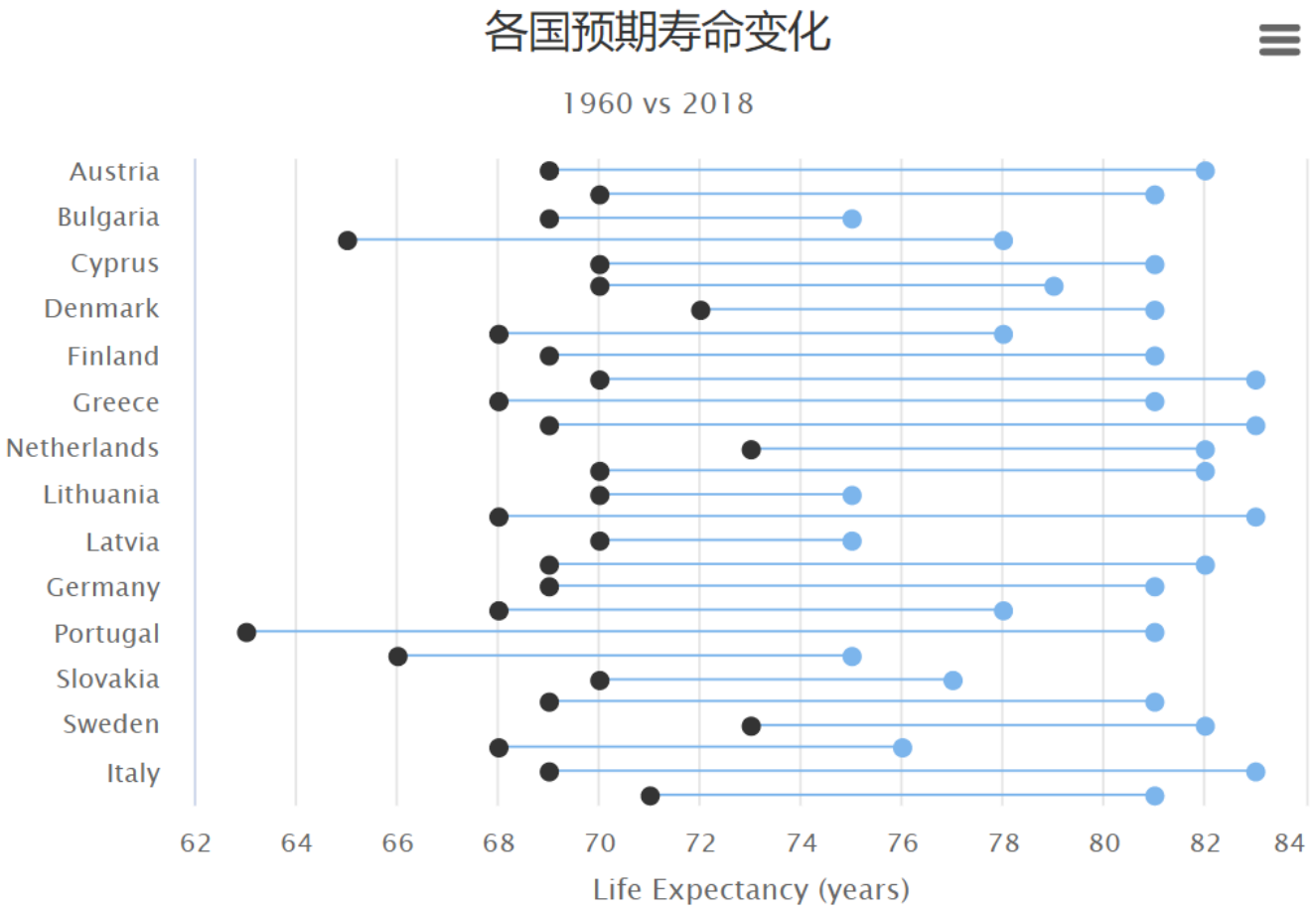
Cleveland点图

- Cleveland点图用点在坐标轴上的位置来表示数据变动的图形。
- 适用数据：定性数据
- 主要功能：与柱状图适用的数据类型表达的含义完全相同，好处在于用点替代柱状图减少混乱。
- 绘图只需要采用geom_point函数：



哑铃图

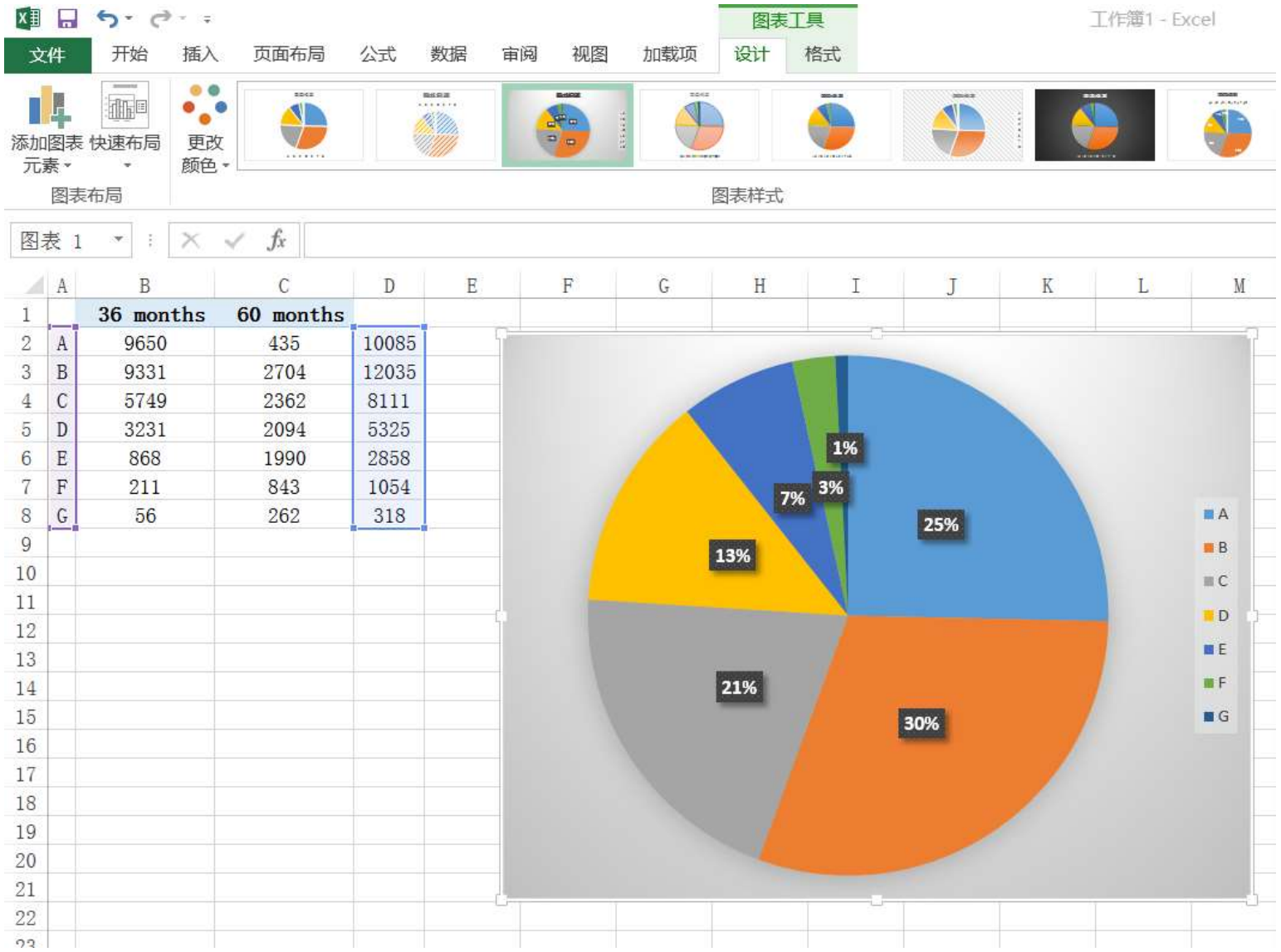
- 比较不同样本或不同分组的某个指标的变化，或者展示不同样本和不同类别之间的关系。
- 使用ggalt包或者plotly包



饼图 (Pie Chart)

- 用圆形及圆内扇形的面积来表示数值大小的图形
- 有单个饼图，也有多主体比较的环形图
- 适用数据：定性数据
- 主要功能：用于展示总体内部的结构，各组成部分所占比例等
- 注意：不能用饼图来展示频数差距！

EXCEL绘制饼图


图表工具
工作簿1 - Excel
文件
开始
插入
页面布局
公式
数据
审阅
视图
加载项
设计
格式
添加图表元素
快速布局
更改颜色
图表布局
图表样式
图表 1
36 months
60 months
A 9650 435 10085
B 9331 2704 12035
C 5749 2362 8111
D 3231 2094 5325
E 868 1990 2858
F 211 843 1054
G 56 262 318
25%
30%
21%
13%
7%
3%
1%
A
B
C
D
E
F
G

环形图(doughnut chart)

- 中间有“空洞”，总体的每一部分数据用环中的一段表示
- 可同时绘制多个总体数据系列，每一总体为一个环
- 主要用于展示两个或多个分类变量的构成比较

R-绘制饼图

基础包中的pie函数也可以绘制饼图

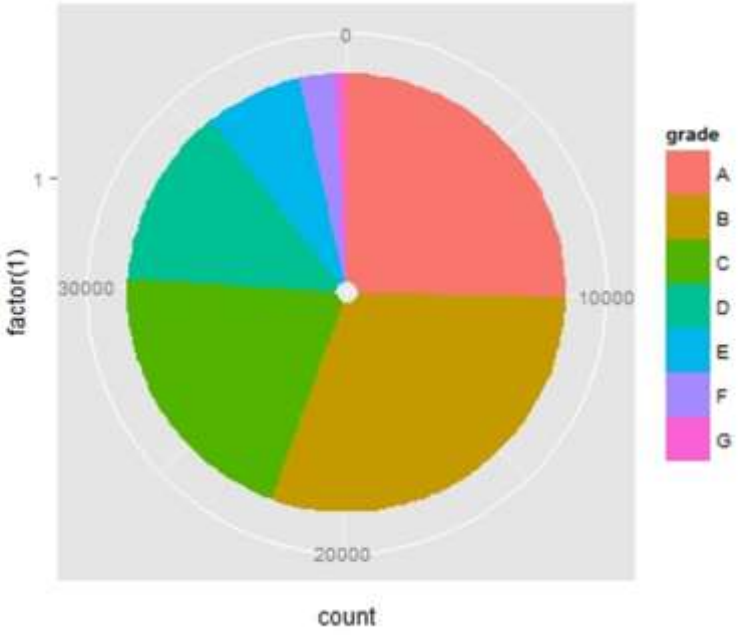
在ggplot2中，饼状图可以视作堆叠的柱状图进行极坐标的变换

以grade变量为例：

```
ggplot(data,aes(x=factor(1),fill=grade))
+geom_bar()+coord_p
```

变化为极

x=factor(1)可以看成设置
之前的y轴被映射为极坐标
也可以看出，在极坐标中，
意义。对于每一个分类所
算在以标签的形式加入到
法是相对繁琐的。

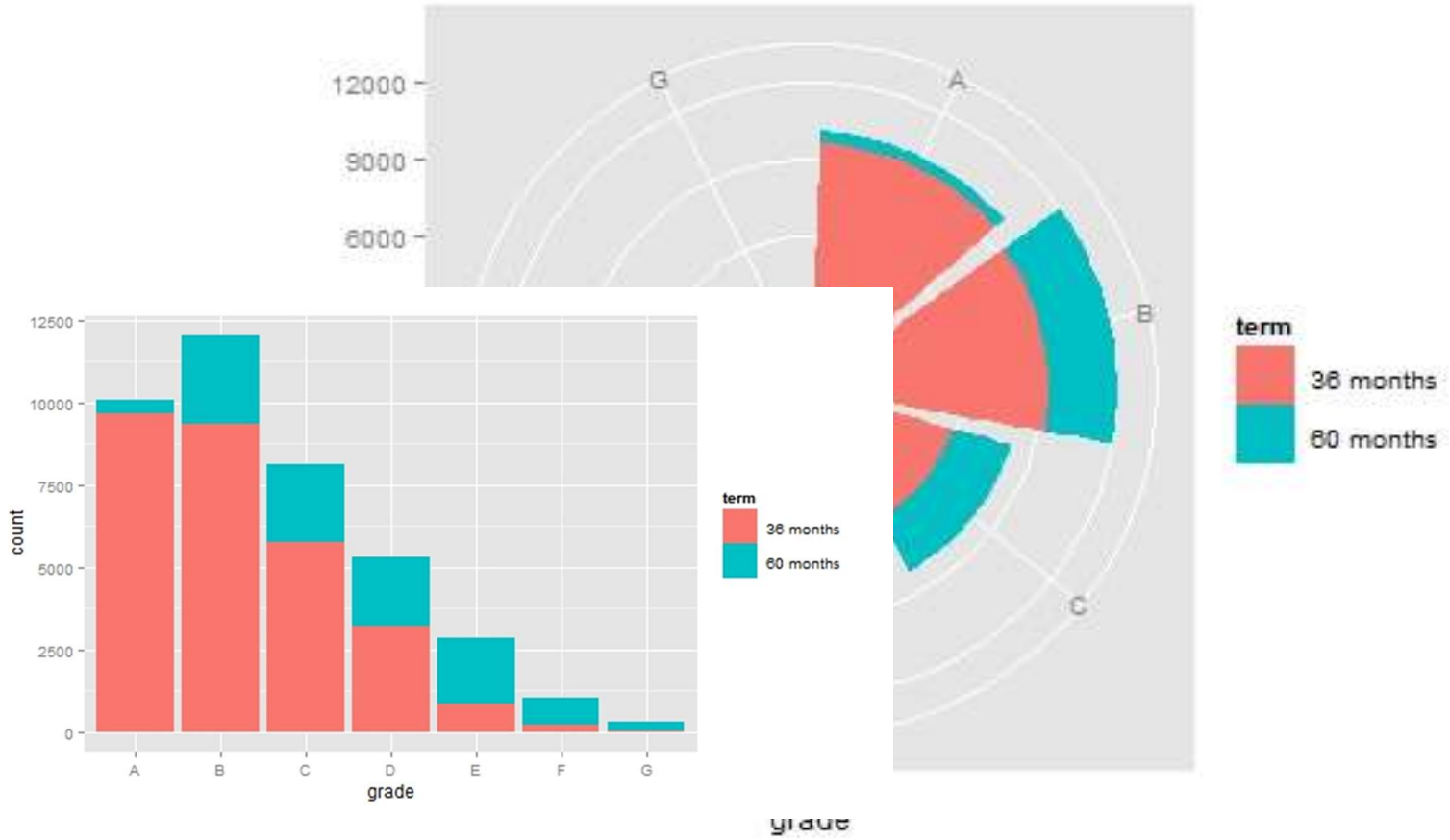


风玫瑰图

- 风玫瑰图是气象科学专业统计图表，用来统计某个地区一段时期内风向、风速发生频率，又分为“风向玫瑰图”和“风速玫瑰图”；因图形似玫瑰花朵，故名。
- 适用数据：定性数据
- 主要功能：考察一个定性变量的频数差异在另一个定性变量分类下的体现

绘制风玫瑰图

- 如果原图是分类变量堆积柱状图，可以用同样的方法绘制风玫瑰图，此时coord_polar（）选择的是默认参数x=“theta”

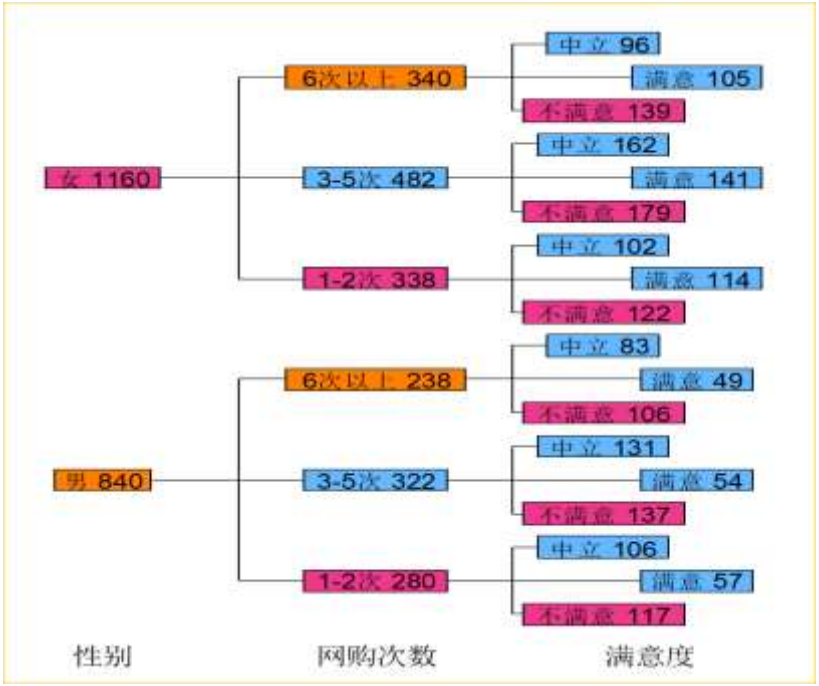


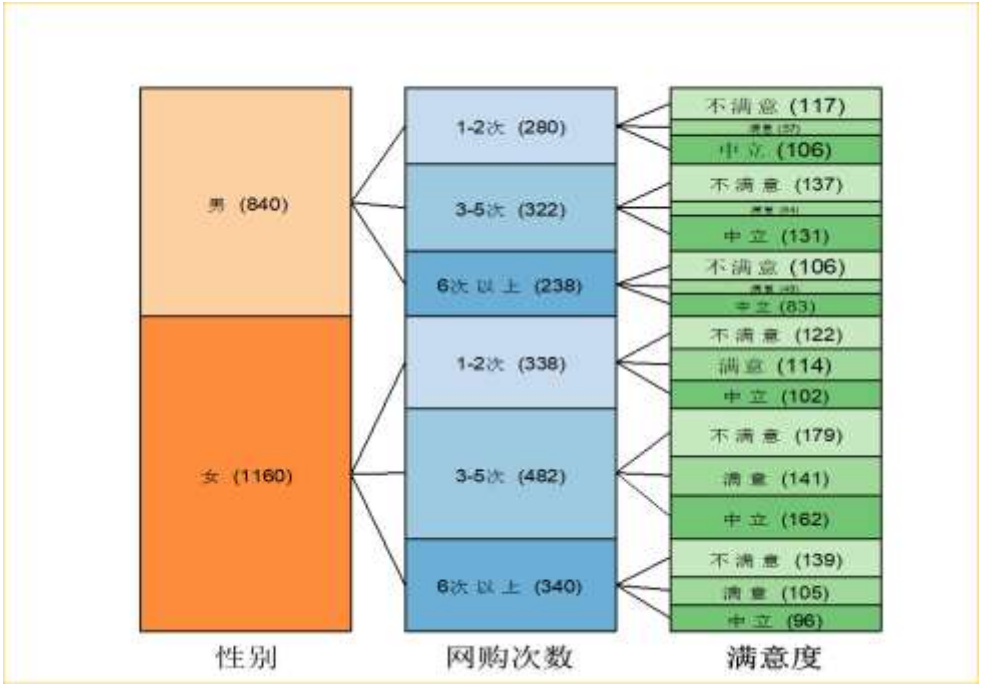
圆形统计图还可以展示什么数据？

来自卫报的交互式可视化作品，其名称为Gay Rights, State by State。同性恋权益在美国这样的联邦国家各州各不相同，为了清晰表达到底哪个州是同性恋的天堂或是地狱，作品以象征同性恋的彩虹入图，十分有创意，不同颜色代表不同权利，如结婚、教育、就业等。这张图使人非常容易了解美国各州的同性恋相关情形：东北部是对同性恋者最友善的地区，之后依序为西南、中西、西北部，而东南部对同性恋者最为苛刻，除了两个州法定允许上课外，其余权益皆无法律保护。而在作者划分的七个权益中，结婚是最少州所允许的。

树状图

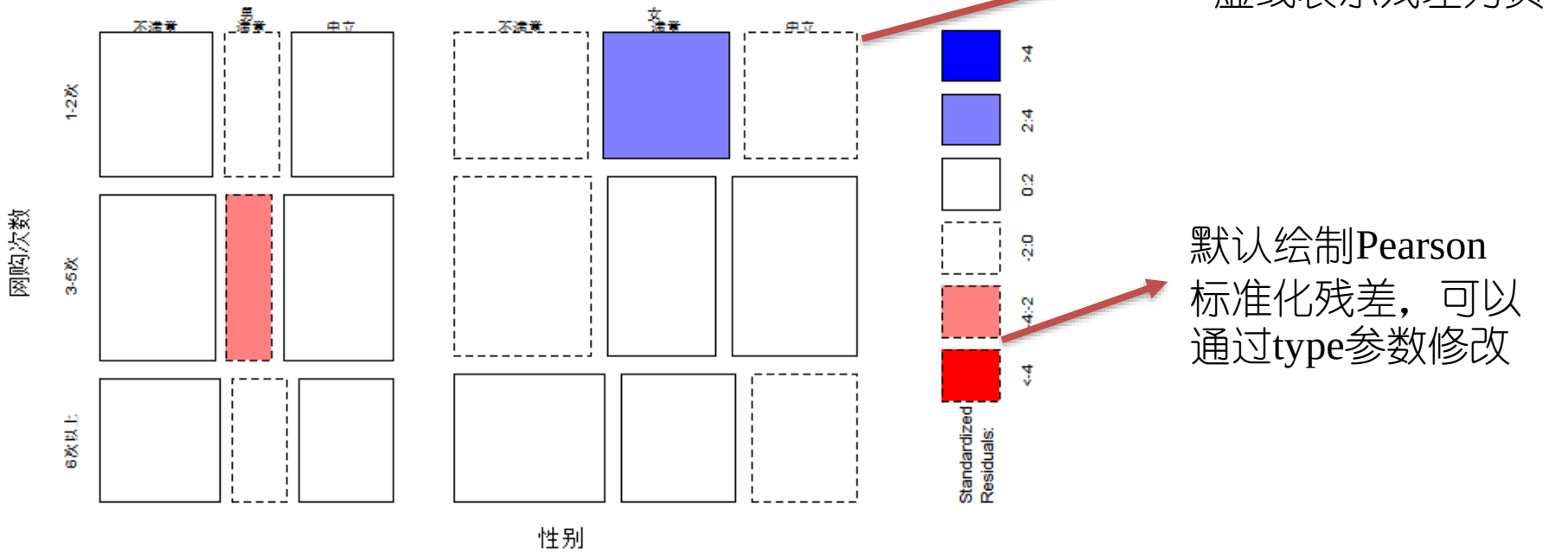
- 将各类别的层次结构画成树状图的形式，称为树状图（dendrogram）或分层树状图
- 使用plotrix包中的plot.dendrite函数和sizetree函数可以绘制





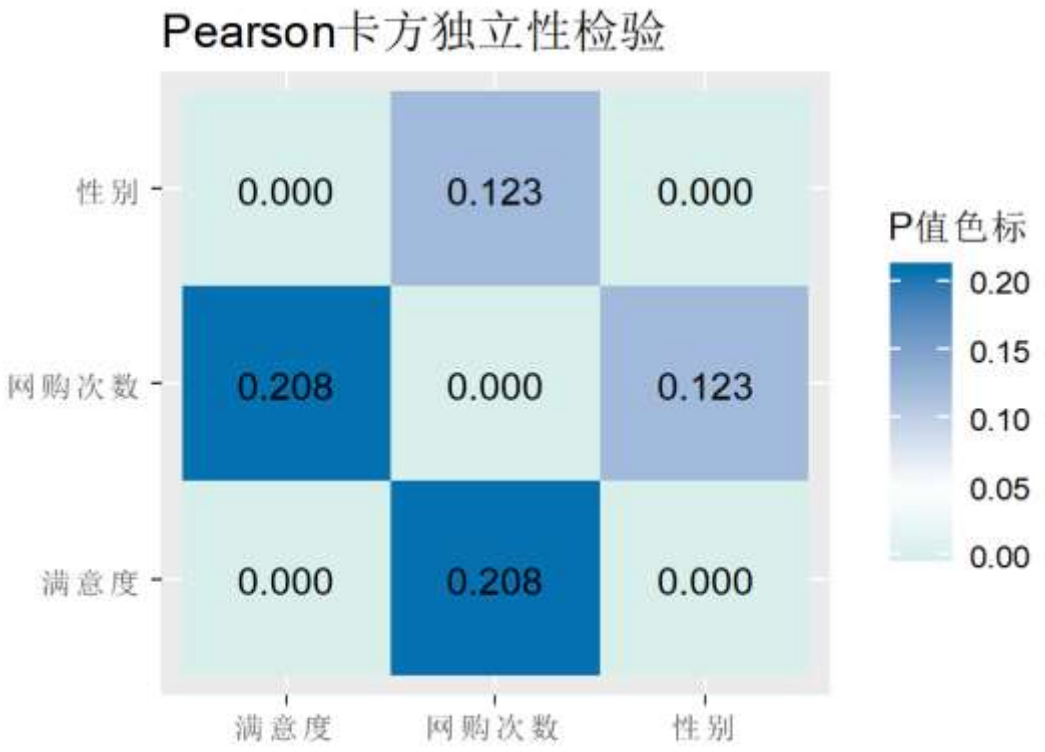
马赛克图

- 马赛克图是高维列联表的可视化图形之一，图中嵌套矩形的面积与列联表相应单元格的频数成比例。
- 使用graphics包中的mosaicplot函数，vcd包中的mosaic、struplot函数都可以创建马赛克图



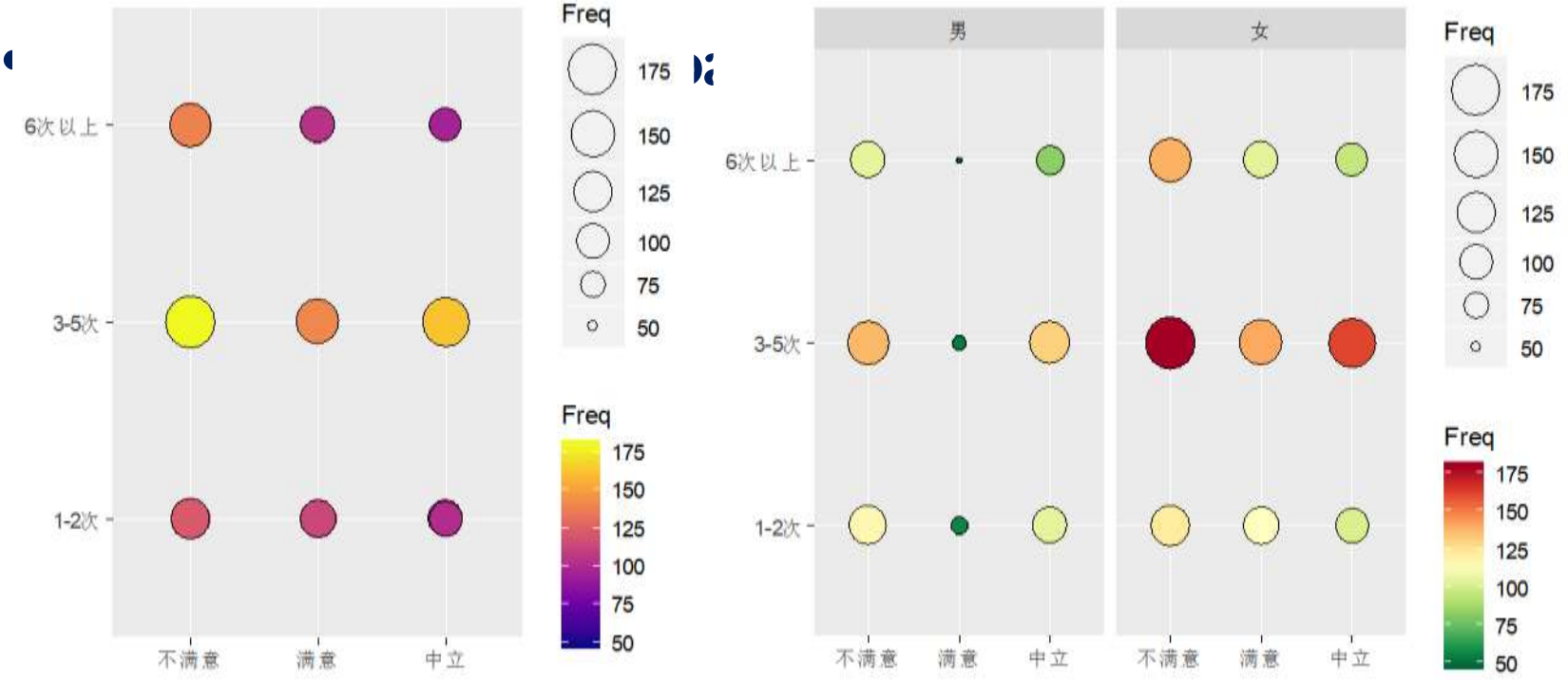
独立性检验的P值图

- 对于多个类别变量，如果要分析任意两个变量之间是否独立，可以使用Pearson卡方检验
- 独立性检验的P值图则列出Pearson卡方检验的P值
- 使用sjPlot包中的sjp.chi2函数可以绘制



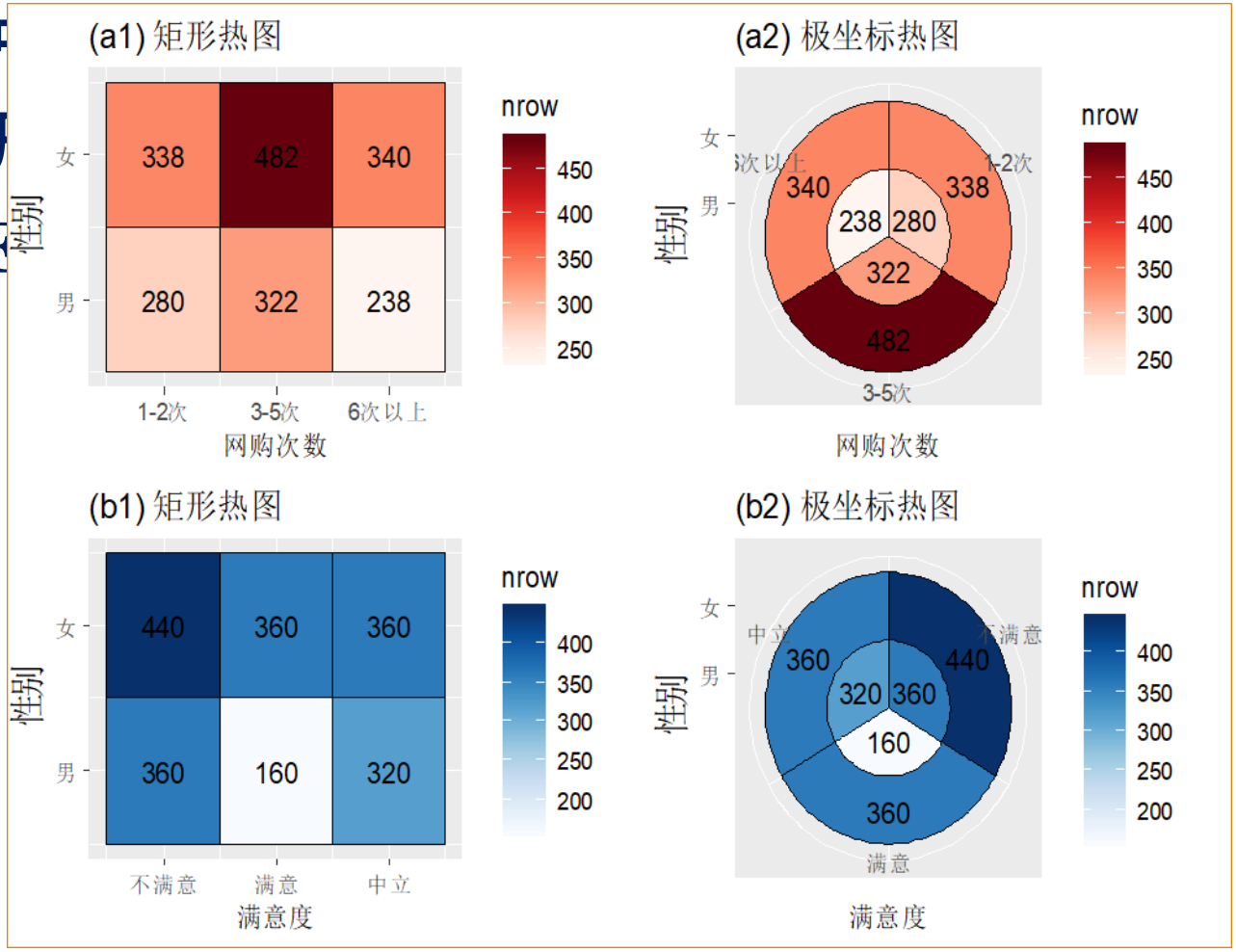
气球图

- 气球图——用气球大小表示数据的图形，它画出的是一个图形矩阵，其中每个单元格包含一个点（气球），其大小与相应数据的大小成比例
- 气球图可用于展示由两个类别变量生成的二维列联表，也可以用于展示具有行名和列名称的其他数据



热图

- 热图是用颜色的饱和度（深浅）表示数值大小图形
- 可以绘制成矩形的形式，用每个矩形的颜色饱和度表示二维表中
- 也可以将矩形
- 使用ggiraphE



词云图(word cloud)

- “词云”这个概念由美国西北大学新闻学副教授、新媒体专业主任里奇·戈登（Rich Gordon）于提出
- 词云是一种可视化描绘单词或词语出现在文本数据中频率的方式，它主要是由随机分布在词云图的单词或词语构成，出现频率较高的单词或词语则会以较大的形式呈现出来，而频率越低的单词或词语则会以较小的形式呈现。
- 适用数据：词频数据
- 主要功能：过滤掉大量的文本信息，使读者只要一眼扫过文本就可以领略文本的主旨
- 在R语言的包里面，有一个wordcloud的包，专门用来创建这种类型的图形，它是由加州大学洛杉矶分校的专业统计学家Ian Fellows编写的。

词云图

wordcloud2(demoFreq)

wordcloud2(demoFreqC, size = 1.55,figPath =log)

letterCloud(demoFreqC,word="R",size = 2)

letterCloud(demoFreqC,word="挖",size = 2)

其他词云工具

- Python：画图的库matplotlib，词云生成库wordcloud和 jieba的分词库
- Wordle：标签云生成工具，可说是这类工具的鼻祖。你只需输入一个网址，就能为这个网页生成关键词标签云。各个关键词的大小与其出现频率成正比。你还可以方便地定制标签云的展现形式。
- BlueMC词云工具
- 图悦picdata.cn/
- BDP个人版https://me.bdp.cn/home.html

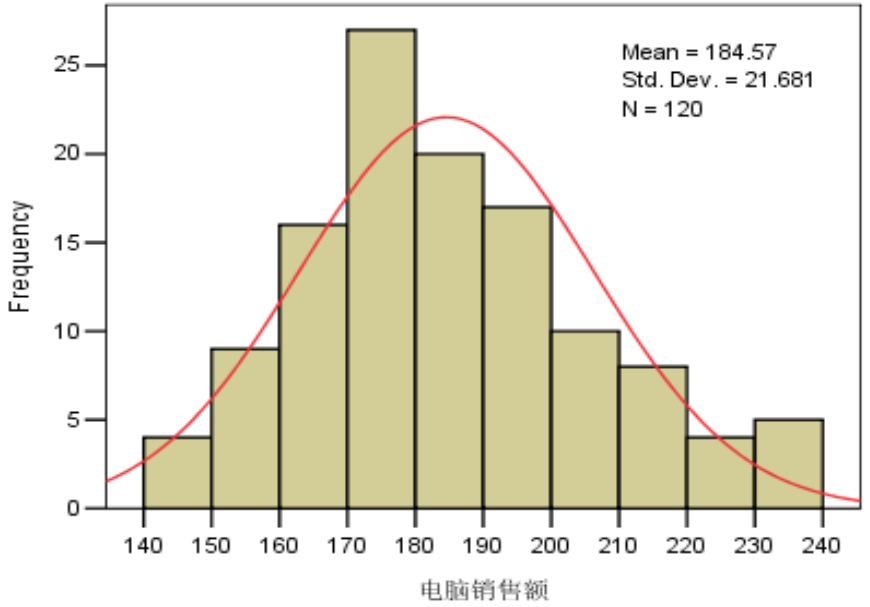
3.3 定量变量数据的展示

定量变量数据的展示

- 直方图
- 密度图
- 峰峦图
- 箱线图
- 小提琴图
- 分布概要图
- ……

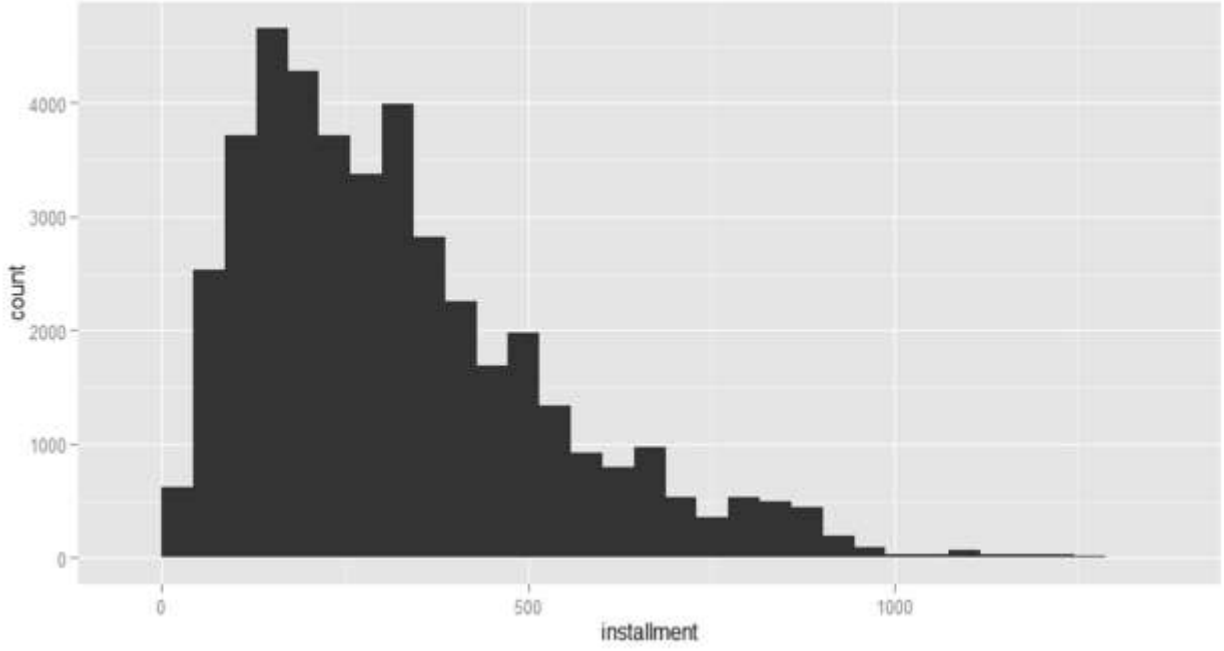
直方图(Histogram)

- 直方图：在统计分组的基础上，用横轴表示数据分组，纵轴表示频数或频率，各组与相应的频数形成一个矩形，即直方图。
- 适用数据：定量数据
- 主要功能：用来反映定量变量的分布状况。



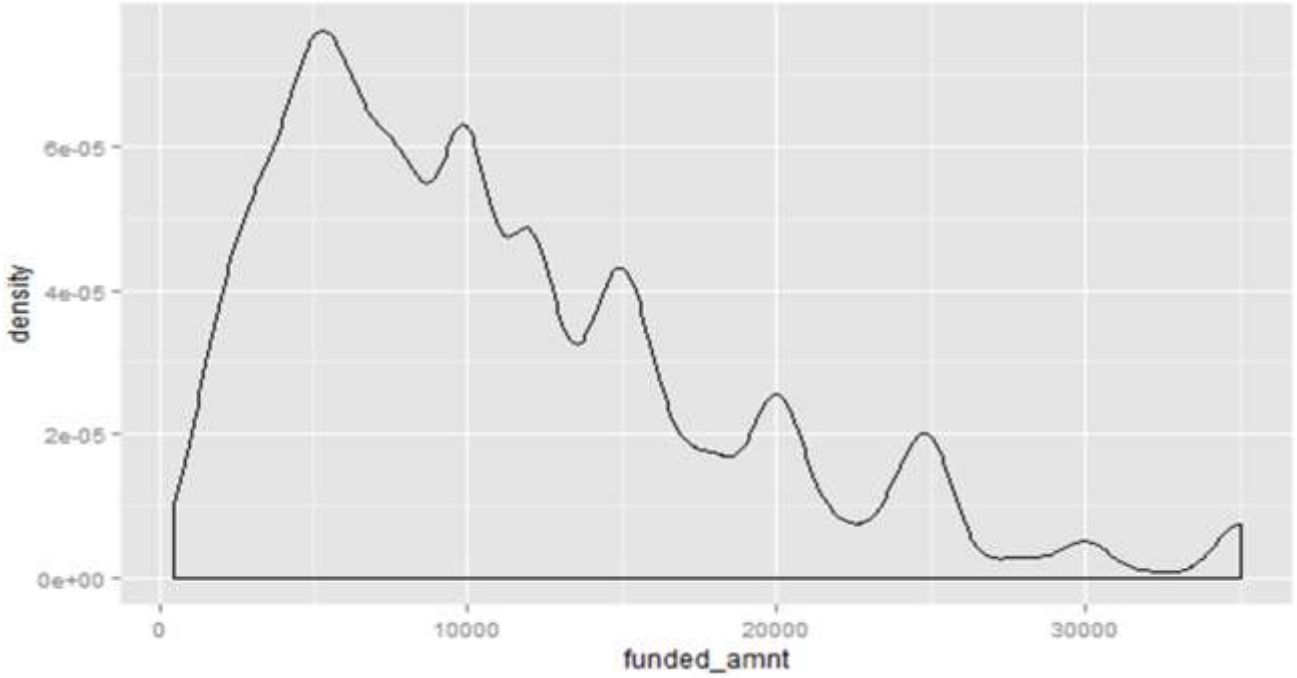
R-频数直方图

- 事实上直方图与柱状图的差别是显著的，他们对应的数据类型就是不同的，但在ggplot2操作中，可以认为只是函数的差异。
- 以原数据集的installment变量为例：
- `geom_histogram`



密度曲线

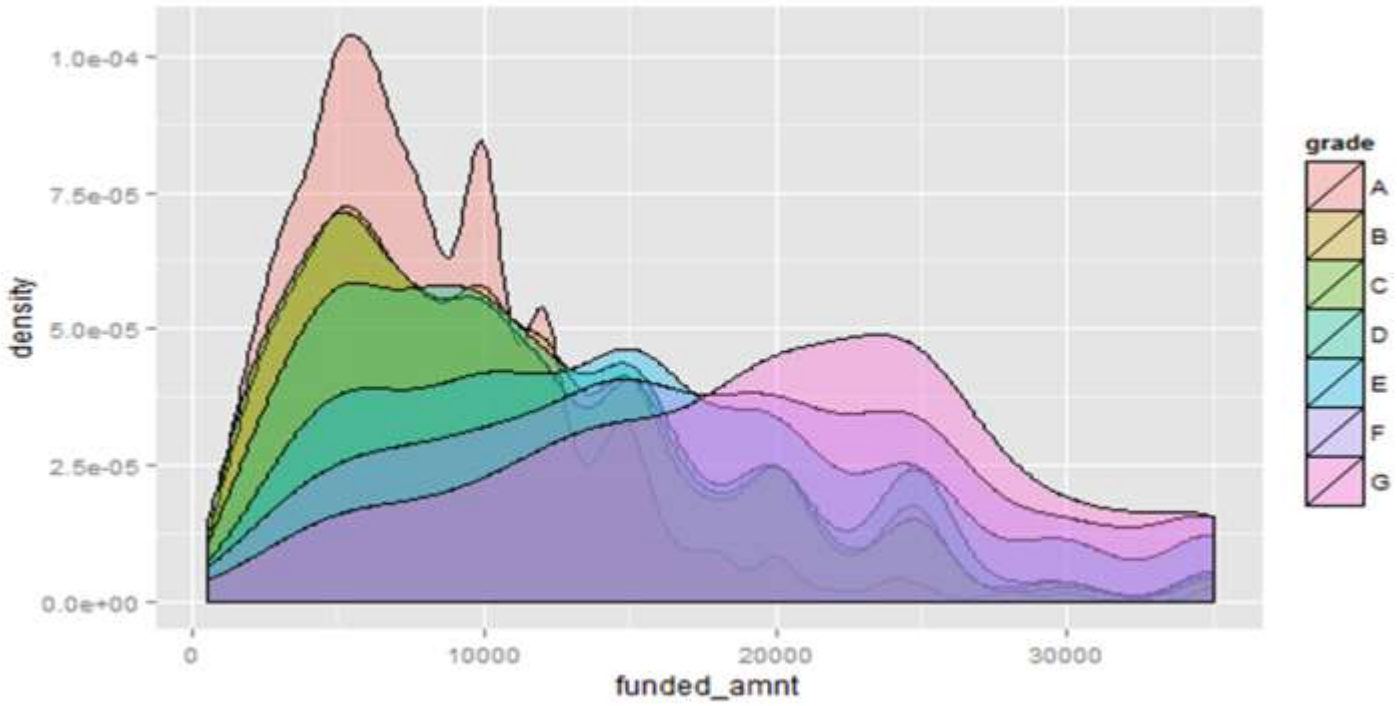
- 核密度估计曲线是基于样本数据对总体分布做出的一个估计。
- 适用数据：定量数据
- 主要功能：它为数值数据的分布提供了一种平滑的描述，从中可以看出分布的大致形状
- 使用geom_



分组密度曲线

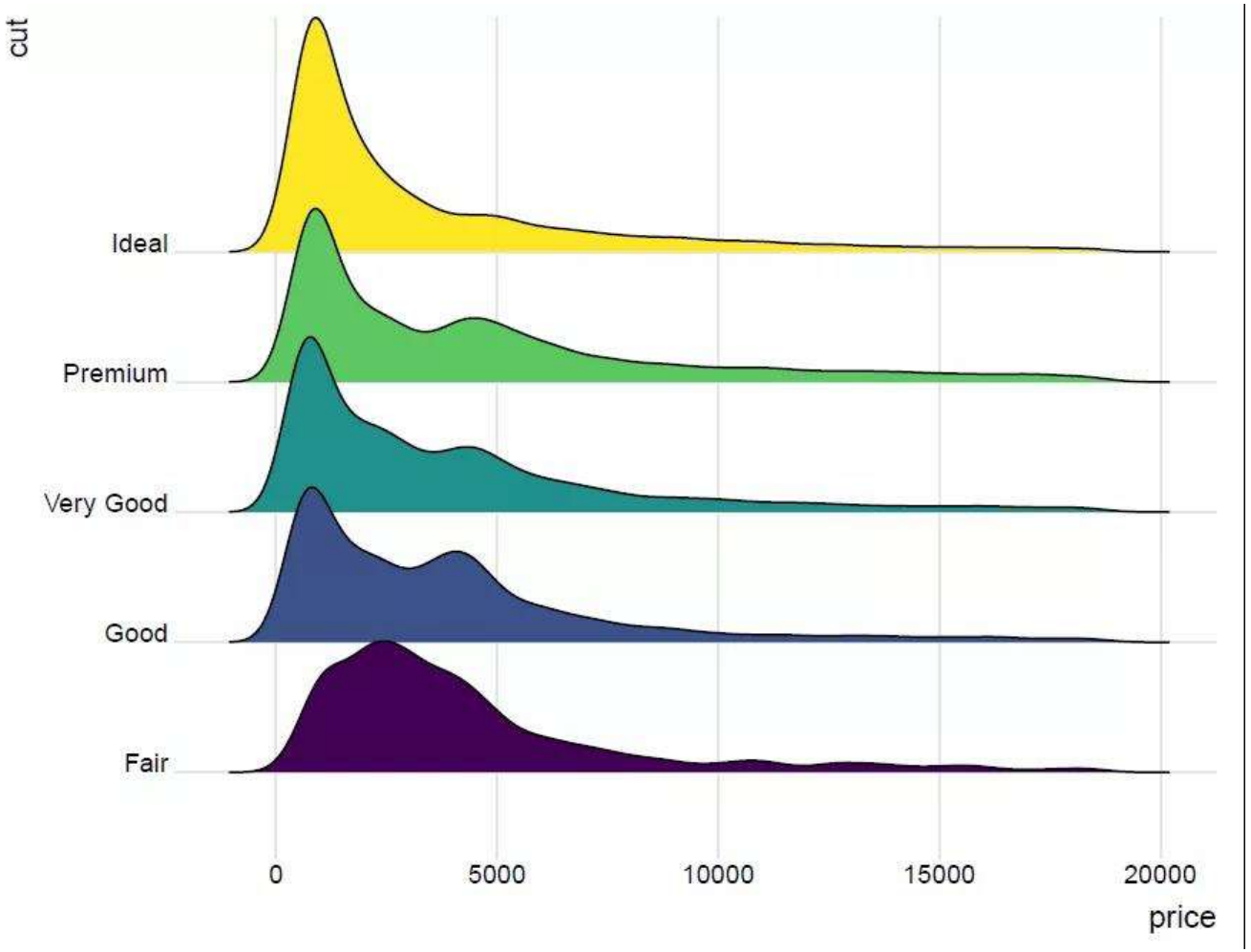
- ggplot(data,aes(x=funded_amnt,fill=grade))+geom_density(alpha=.35)

用alpha调整透明度



峰峦图

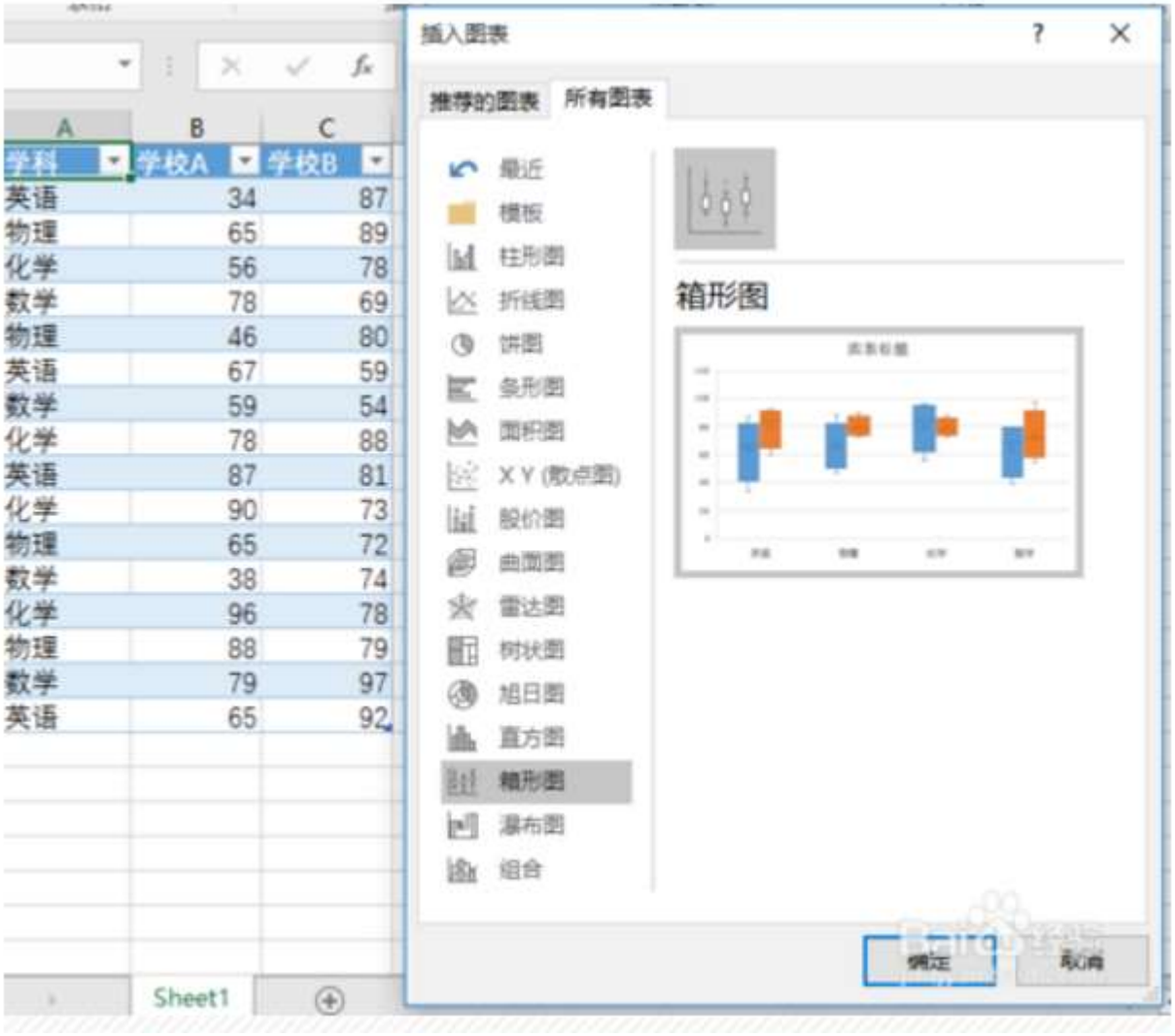
- ggridges包中的geom_density_ridges()



箱线图(Box Plot)

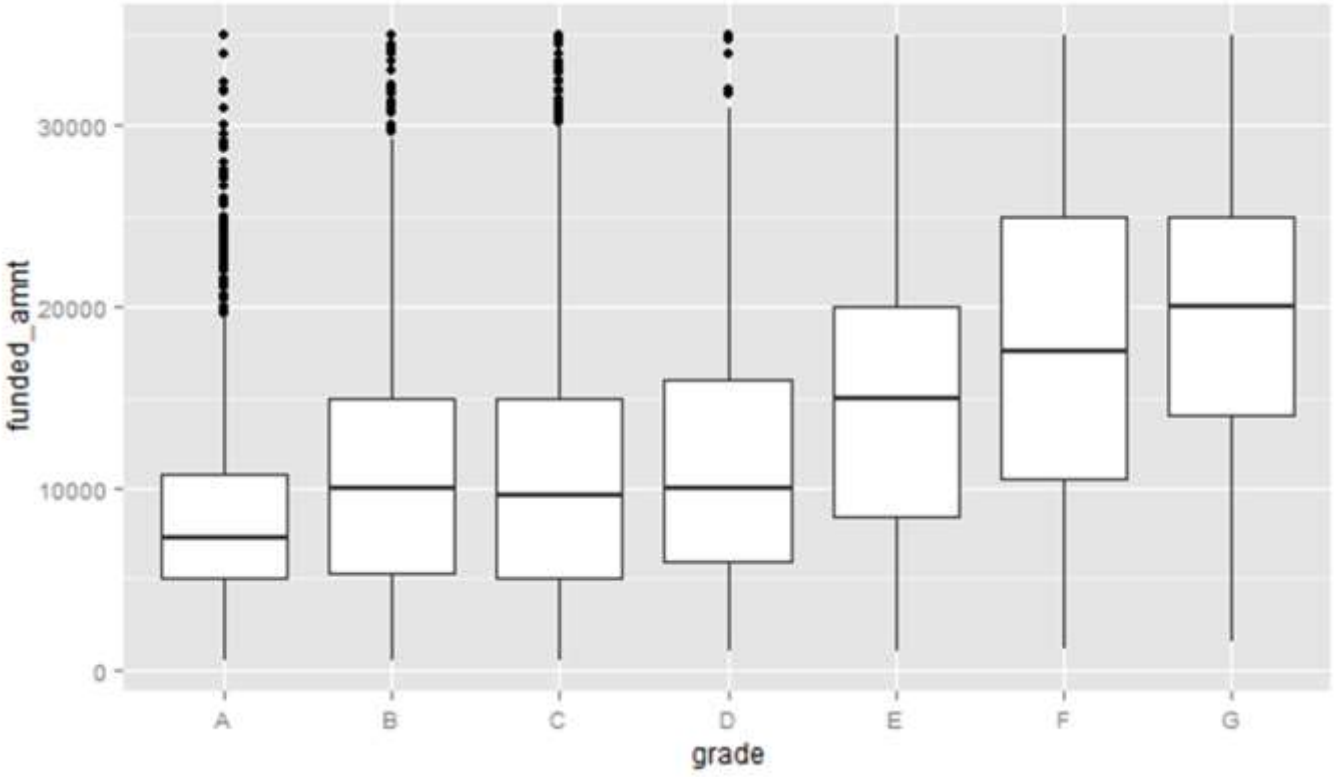
- 箱线图由一组数据的5个特征值绘制而成，一个箱子和两条线段组成
- 适用数据：定量数据
- 主要功能：用于显示未分组原始数据的分布，它不仅可以反映一组数据分布的特征如分布是否对称，是否存在离群点等，还能够进行多组数据分布特征的比较。

EXCEL-箱线图


A
B
C
学科
学校A
学校B
英语 34 87
物理 65 89
化学 56 78
数学 78 69
物理 46 80
英语 67 59
数学 59 54
化学 78 88
英语 87 81
化学 90 73
物理 65 72
数学 38 74
化学 96 78
物理 88 79
数学 79 97
英语 65 92
Sheet1
插入图表
推荐的图表
所有图表
最近
模板
柱形图
折线图
饼图
条形图
面积图
XY (散点图)
股价图
曲面图
雷达图
树状图
旭日图
直方图
箱形图
瀑布图
组合
箱形图
确定
取消

R-箱线图

- 用ggplot来绘制箱线图
- ggplot(data,aes(x=grade,y=funded_amnt))+geom_boxplot()

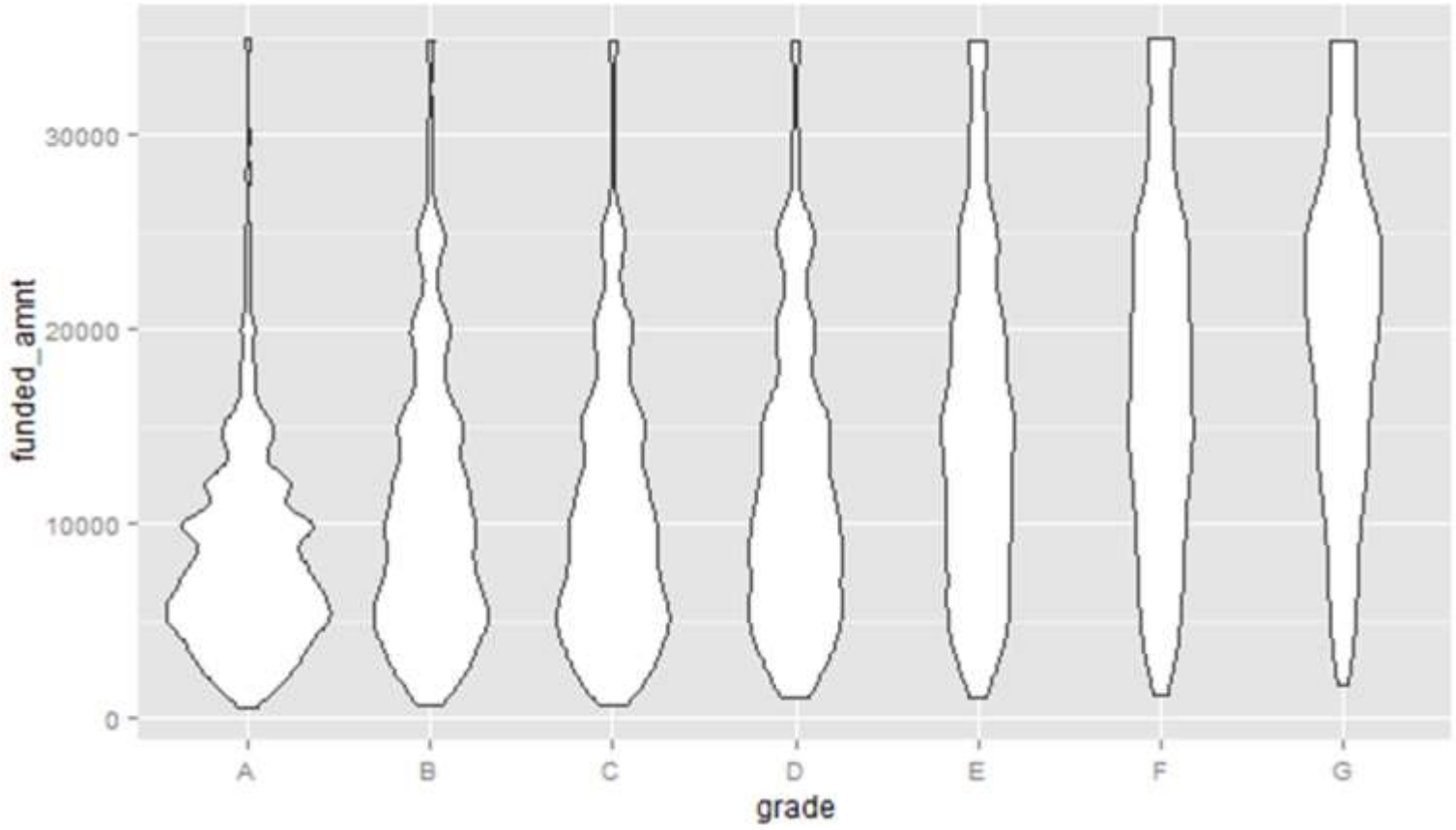


小提琴图

- 小提琴图是一种用来对多组数据的分布进行比较的方法。小提琴图也是核密度估计，但绘图时对核密度曲线取了镜像以使形状对称。
- 适用数据：定量数据
- 主要功能：它为数值数据的分布提供了一种平滑的描述，从中可以看出分布的大致形状。主要用于多主体的比较

小提琴图

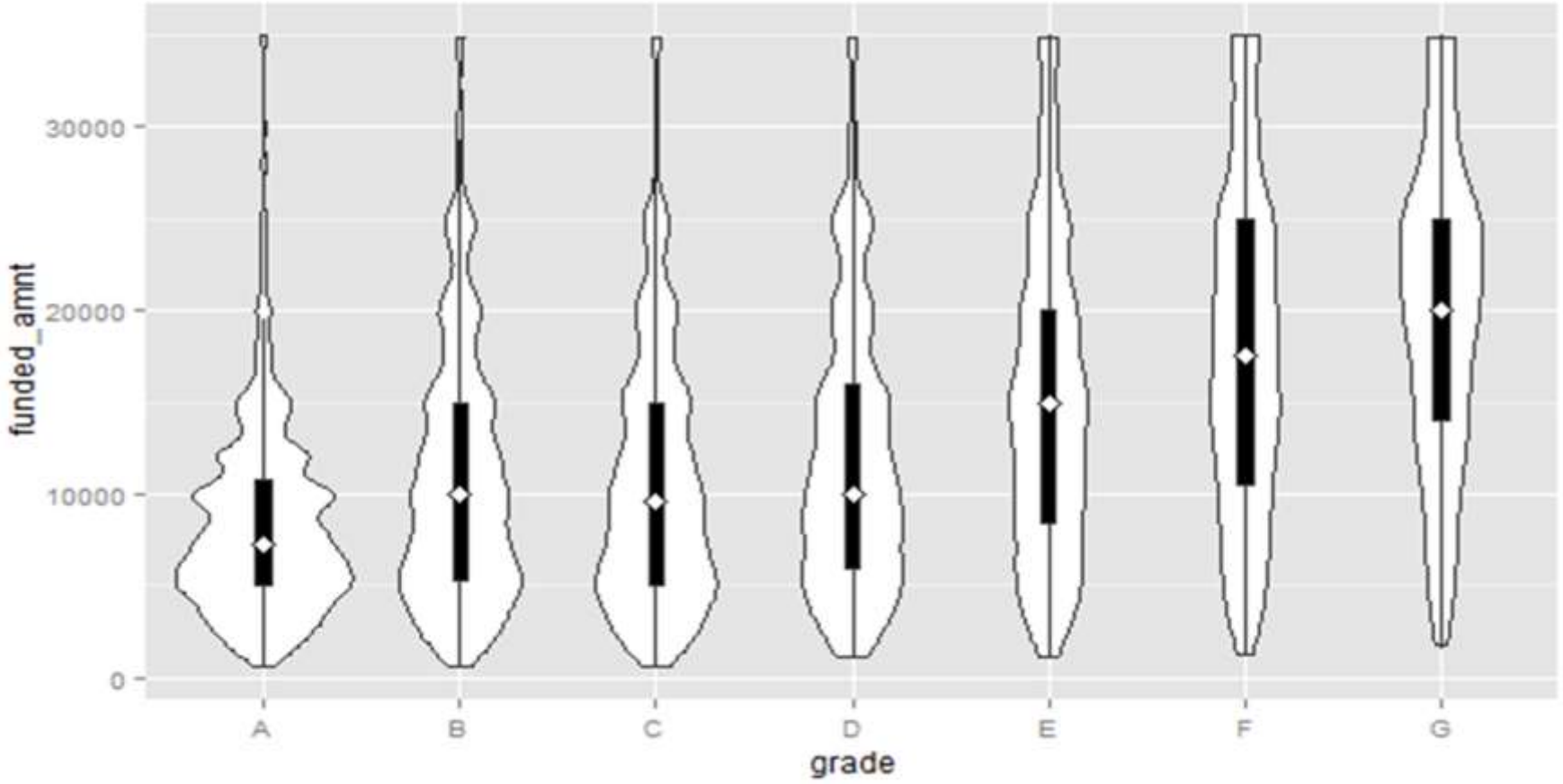
- ggplot(data,aes(x=grade,y=funded_amnt))+geom_violin()



小提琴图

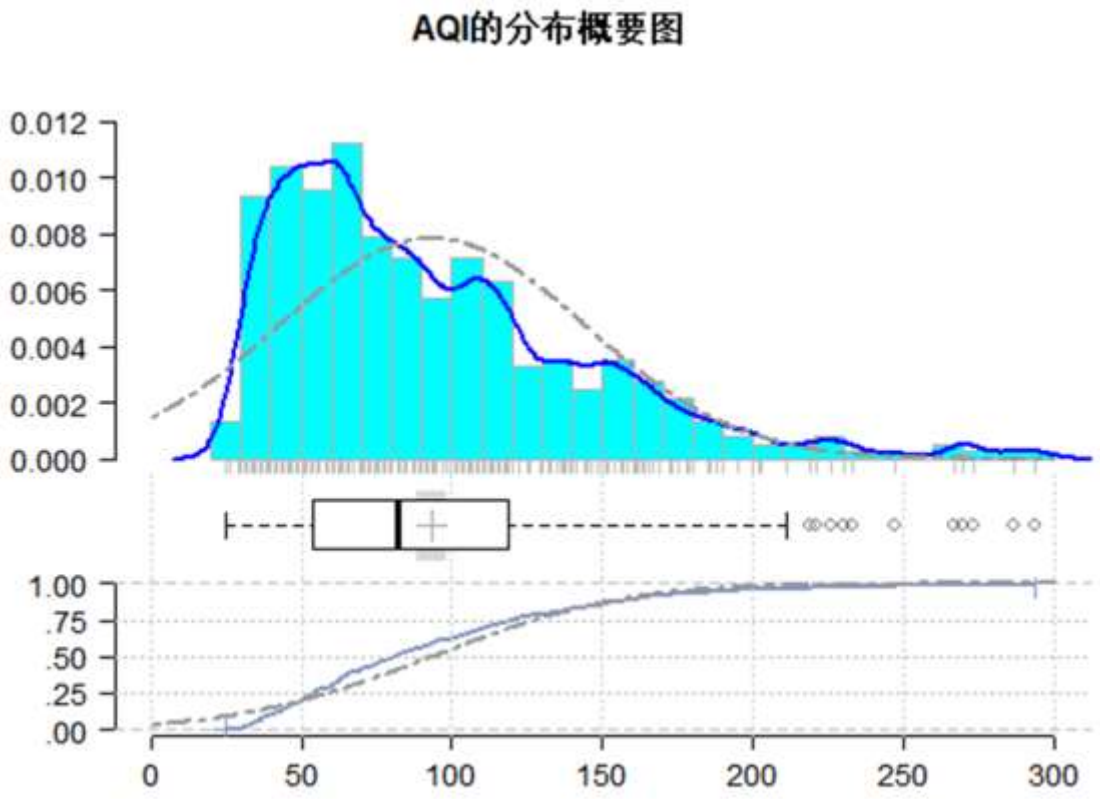
- 可以在小提琴图中叠加箱线图，同时用一个白圆圈表示中位数，通过设置outlier.colour=NA可以隐去箱线图中的异常点。
- 在这个例子中，我们从下而上逐层绘制图形，即先绘制小提琴图，再叠加箱线图，之后使用stat_summary()计算并绘制表示中位数的白圆圈。
- ggplot(data,aes(x=grade,y=funded_amnt))
- +geom_violin()
- +geom_boxplot(width=.1,fill="black",outlier.colour=NA)
- +stat_summary(fun.y=median,geom="point",shape=23,size=3,fill="white")

小提琴图



分布概要图

- 如果想用一幅图对数据的分布特征有一个概括性的描述，可以使用aplpack包中的plotsummary函数和DescTools包中的PlotFdist。只分析一个变量时，可以用用PlotFdi
数将直方图、
组合在一个图
线等叠加在图

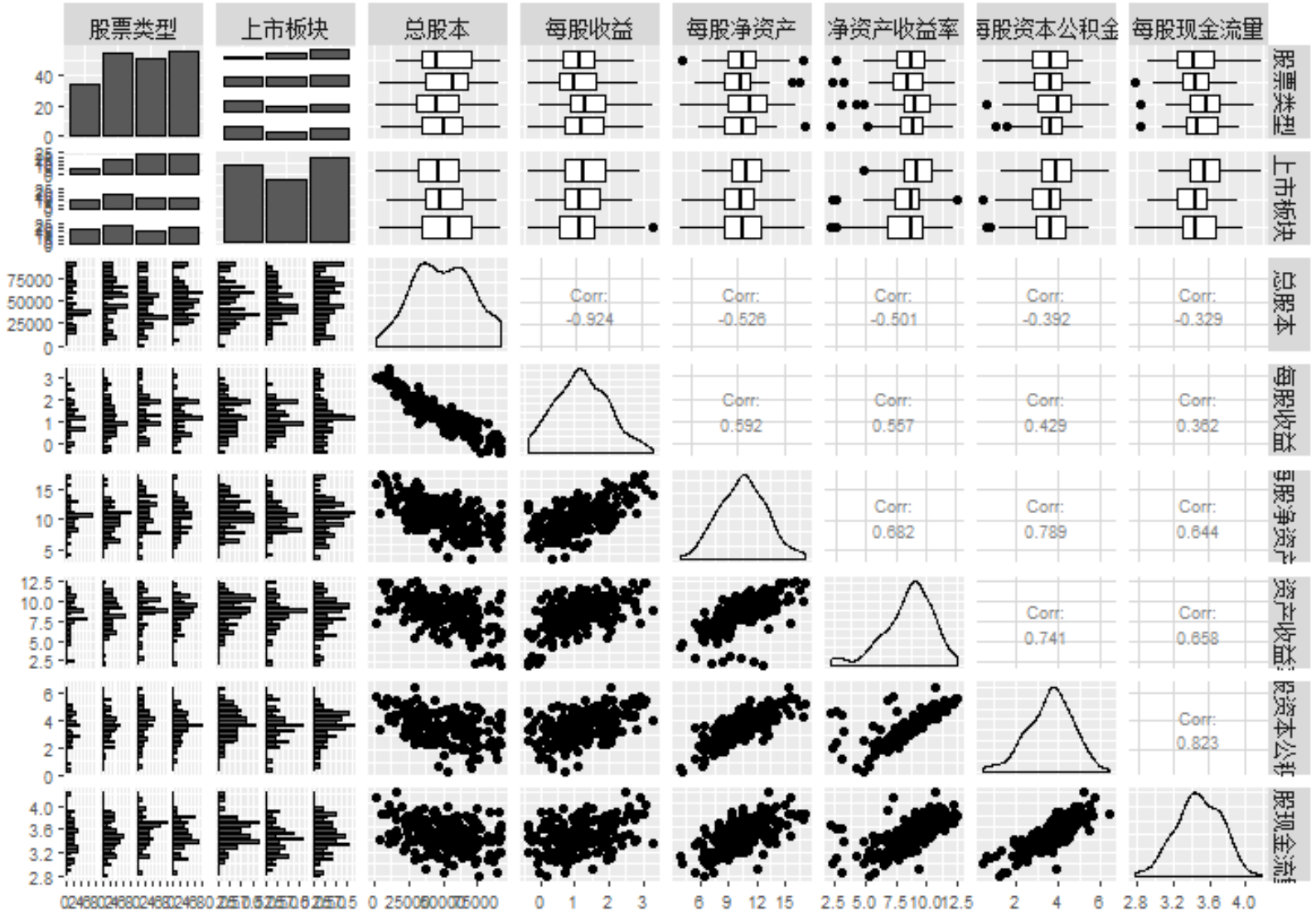


广义配对图阵

- 当数据集中既包含类别变量，又包含数值变量时，可以根据变量类型分别绘制不同的图形以及按类别变量分类的数值变量的图形，以展示复杂数据集和类别变量之间的对等关系。这样的图形就是广义配对图。
- 使用GGally包中的ggpairs函数和gpairs包中的gpairs函数均可以绘制广义配对图。函数会根据给定的变量绘制一个图形矩阵，矩阵中的图形取决于变量的类型。

广义配对图阵

- GGally::ggpairs(data5_1)



3.3 变量之间的关系的展示

变量之间的关系的展示

- 散点图
- 密度图
- 气泡图
- 折线图
- 广义配对图
- ……

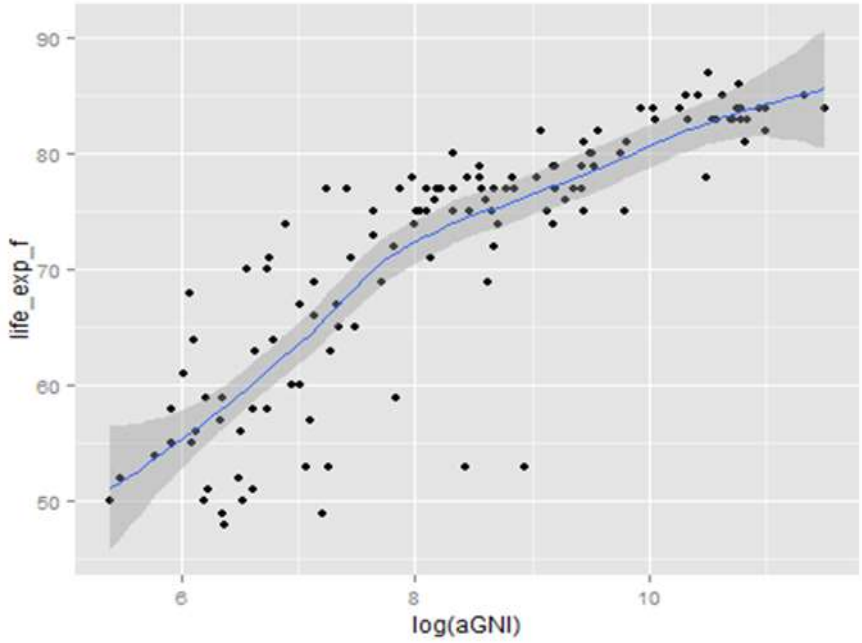
散点图(Scatter plot)

- 散点图：每组数据(xi , yi)在坐标系中用一个点表示，n组数据在坐标系中形成的n个点称为散点，由坐标及其散点形成的图
- 有2维、3维、矩阵等形式
- 适用数据：两个定量数据
- 主要功能：用点在坐标系上的位置展示两个数值变量的分布，探索它们间的相关关系。

散点图

- **散点图中，可以添加各种类型的回归趋势线直观地展示两个变量的关系。**

```
#创建一个图层对象
p<-ggplot(keyindicators,aes(x=log(aGNI),y=life_exp_f))+geom_point()
#添加一条线性拟合曲线，置信度默认为0.95
p+stat_smooth(method=lm)
#改置信度为0.99
p+stat_smooth(method=lm,level=0.99)
#不添加置信区间
p+stat_smooth(method=lm,se=FALSE)
#添加局部加权多项式回归线
p+stat_smooth(method=loess)
```

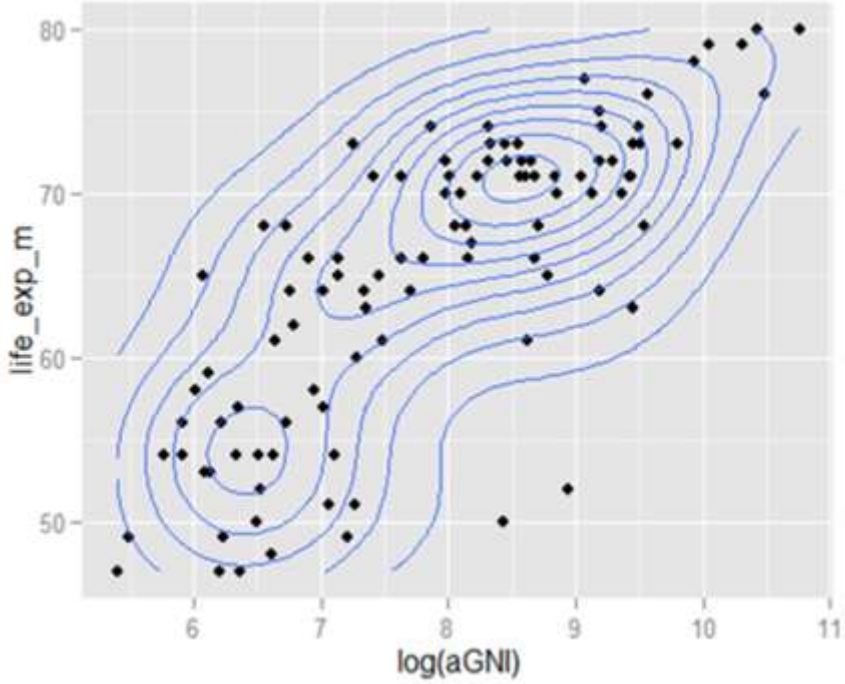


二维变量的密度图

- 地理学里，地图上地势高度相同的点连成的曲线被称之为等高线，将地势高度转化为二元随机变量联合密度函数的取值，那么等高线图就能借以绘制一个二维随机变量的核密度估计图。
- 适用数据：两个定量数据
- 主要功能：展示二维随机变量的密度分布情况。

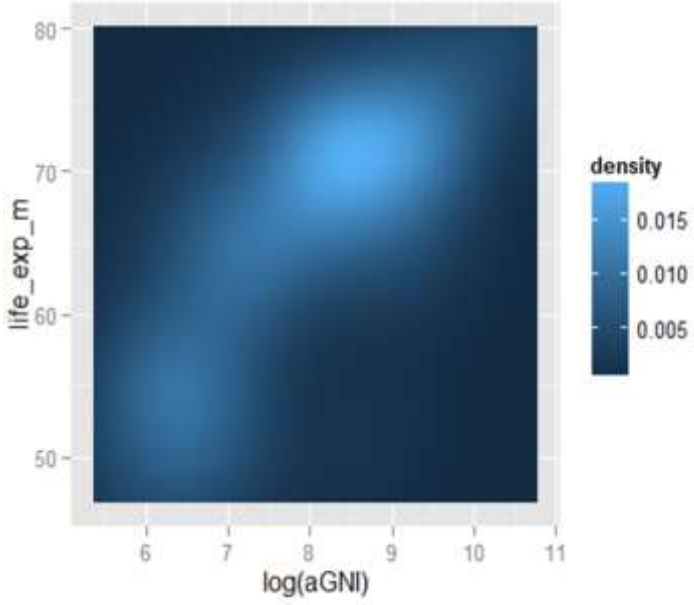
密度图

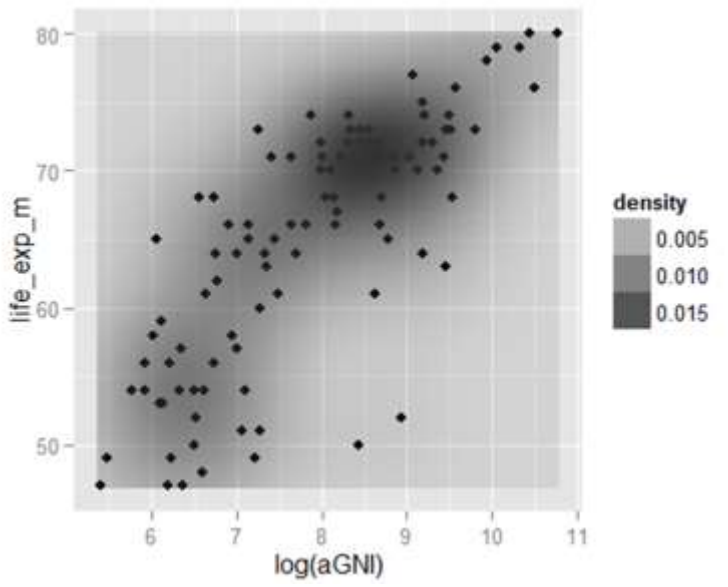
- 二维核密度估计图的展示方式有多种，在ggplot2中它们都是用stat_density()函数来实现的
- #生成几何对象
- p<-ggplot(keyindicators1,aes(x=log(aGNI),y=life_exp_m))
- #默认等高线图
- p+geom_point()+stat_density2d()



密度图

- 我们既可以得到系统默认的等高线图，也可以通过将density映射给fill或者是alpha来输出填充了颜色的瓦片图。其中，某点的核密度估计使用颜色来显示。
- #有填充颜色的等高线
- p+stat_density2d(aes(fill=..density..),geom = "tile",contour= FALSE)
- #有数据点，并将核密度估计映射给alpha
- p+geom_point()+stat_density2d(aes(alpha=..density..),geom="tile",contour= FALSE)





三个变量

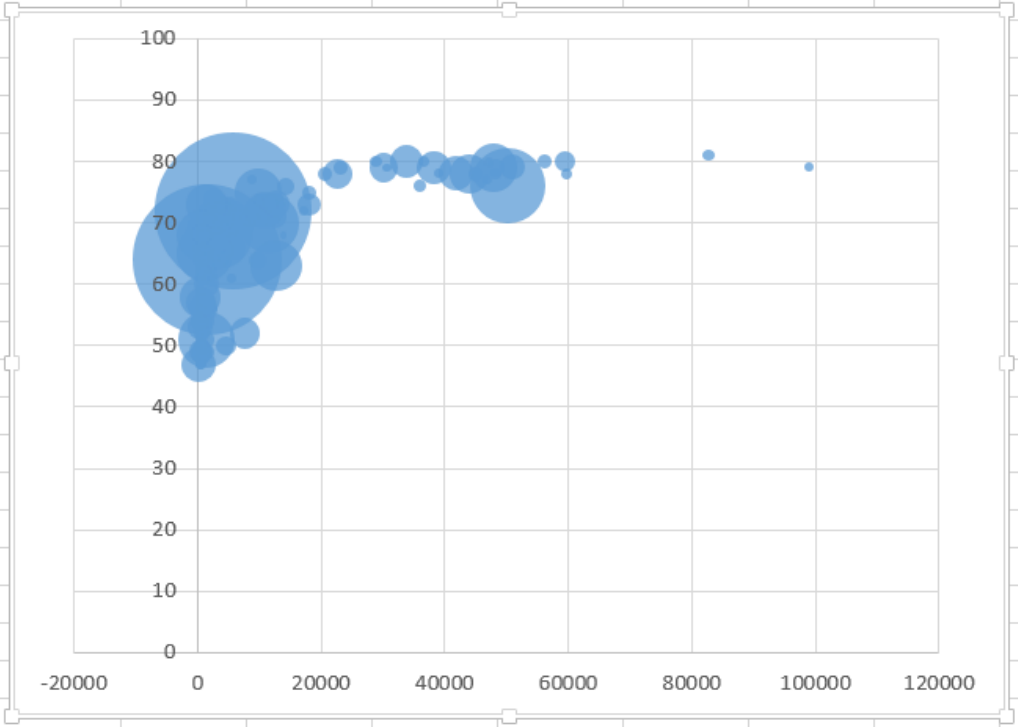
- 三个变量的处理方式：
- 思考：
- 1.直观展示不同地区或者不同收入水平国家的GNI和预期寿命的差别？
- 2.直观展示不同国家人口、国民人均收入、预期寿命的关系？

气泡图

- 气泡图与散点图的适用场景类似，可以理解气泡图是散点图的升级版；不同之处在于气泡图允许在图表中额外加入一个表示大小的变量，组成3个数值进行对比，从而在图表中获取更多的信息。
- 适用数据：三个定量变量
- 主要功能：展示三个变量之间的关系

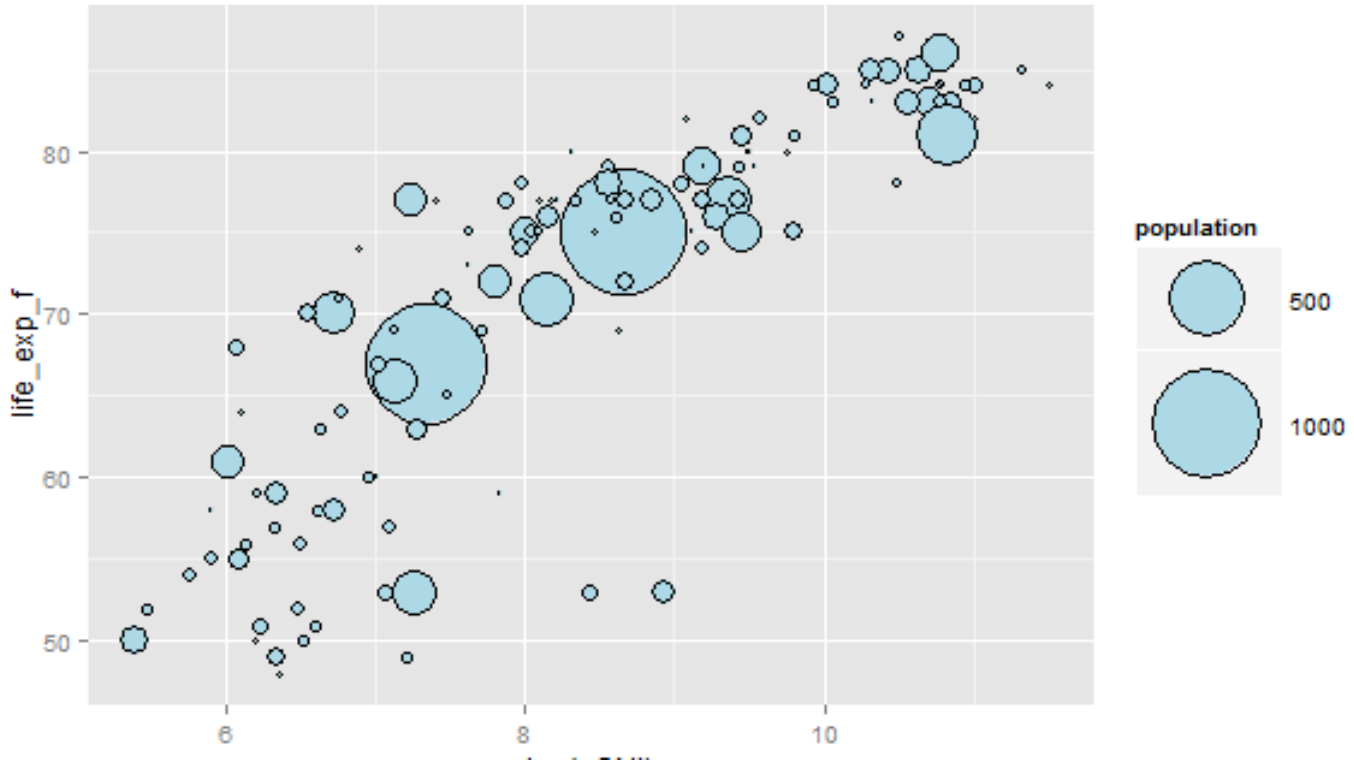
EXCEL-气泡图

	A	B	C	D
1	country	aGNI	life_exp	population
2	Albania	4090	74	3
3	Algeria	4110	72	38
4	Angola	4580	50	21
5	Azerbaij	6050	68	9
6	Belarus	6530	65	9
7	Bosnia a	4650	73	4
8	Brazil	11630	70	199
9	Bulgaria	6870	71	7
10	China	5740	72	1351
11	Colombia	6990	70	48
12	Costa Ri	8740	77	5
13	Dominica	5470	71	10
14	Ecuador	5190	73	15
15	Hungary	12390	71	10
16	Iraq	5870	66	33
17	Jordan	4720	72	6
18	Kazakhst	9730	64	17
19	Lebanon	9190	70	4
20	Malaysia	9800	72	29
21	Mexico	9740	75	121
22	Panama	9910	74	4
23	Peru	5880	71	30
24	Romania	8420	71	21
25	Serbia	5280	72	7
26	South Af	7610	52	51
27	Thailand	5210	71	67
28	Tunisia	4150	73	11



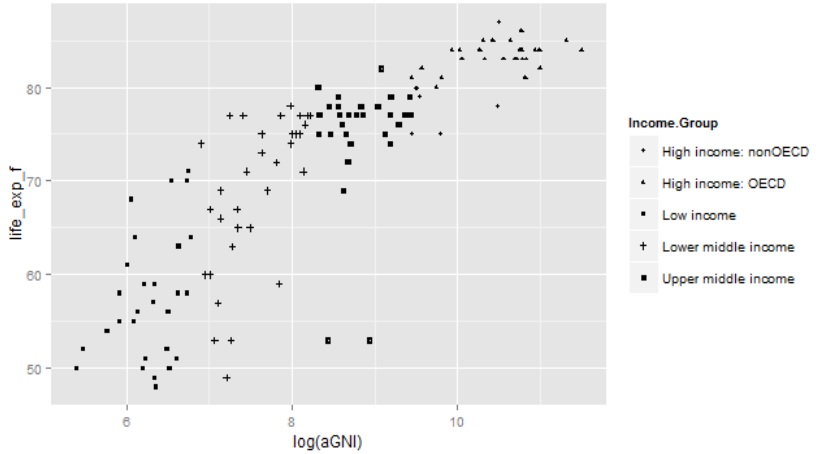
气泡图

- #添加国家人口为第三个变量，展示变量间关系
- ggplot(keyindicators,aes(x=log(aGNI),y=life_exp_f,size=population))+
- geom_point(shape=21,colour="black",fill="lightblue")+
- scale_size_area(max_size = 25)

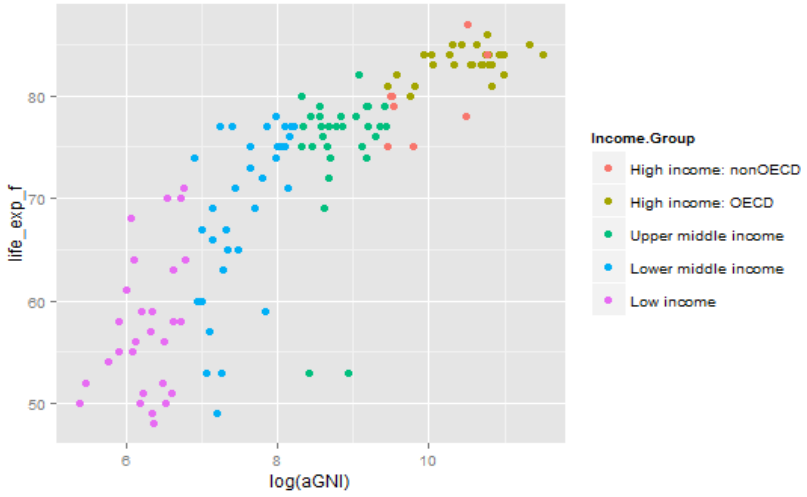


映射到颜色或形状

- 将第三个变量映射到颜色、形状等图形属性上
- #形状
- ggplot(keyindicators,aes(x=log(aGNI),y=life_exp_f,shape=Income.Group))+geom_point(size=1.5)



- #颜色
- ggplot(keyindicato me.Group))+geom_point(size=1.5)



图形分面

- 分面即在一个页面依据一个或几个分类变量，自动摆放多幅图形的技法：现将数据依据分类变量划分为子集，然后将每个子集数据分别绘制到页面的不同面板中，这类图形也通常被称作small multiples。
- 1.facet_grid
- 2.facet_wrap

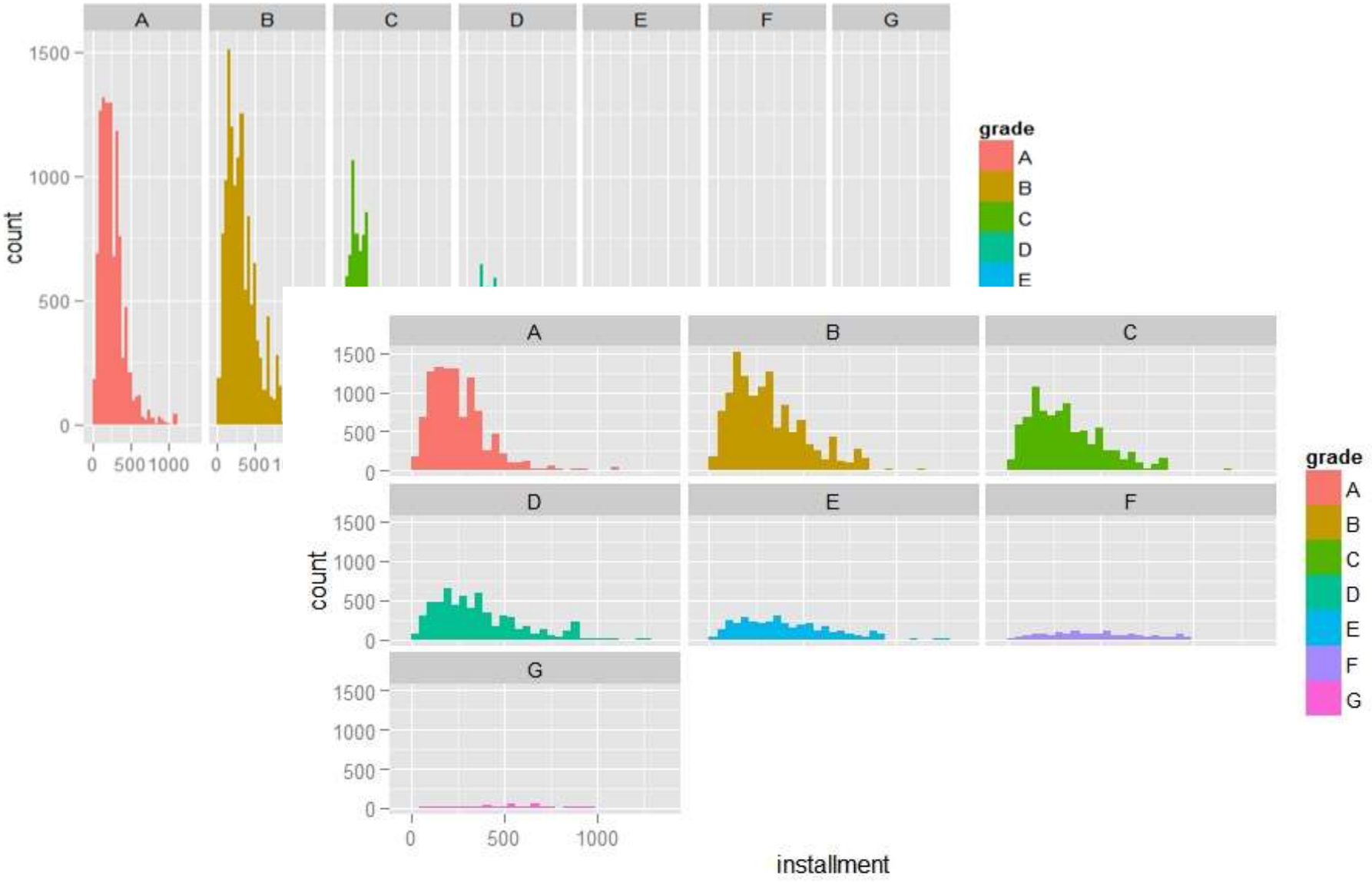
图形分面

- 以grade变量为分类变量：
- 1. `p+geom_histogram()+facet_grid(.~grade)`
- 可以根据数据的不同分组, 将图形按照 水平或者垂直方向 进行分割.共享x轴或者轴.

本应该是两个变量，用“.”省略一个

- 2. `p+geom_histogram()+facet_wrap(~grade)`
- 对数据分类只能应用一个标准，不同组数据获得的小形按从左到右从上到下的顺序进行排列

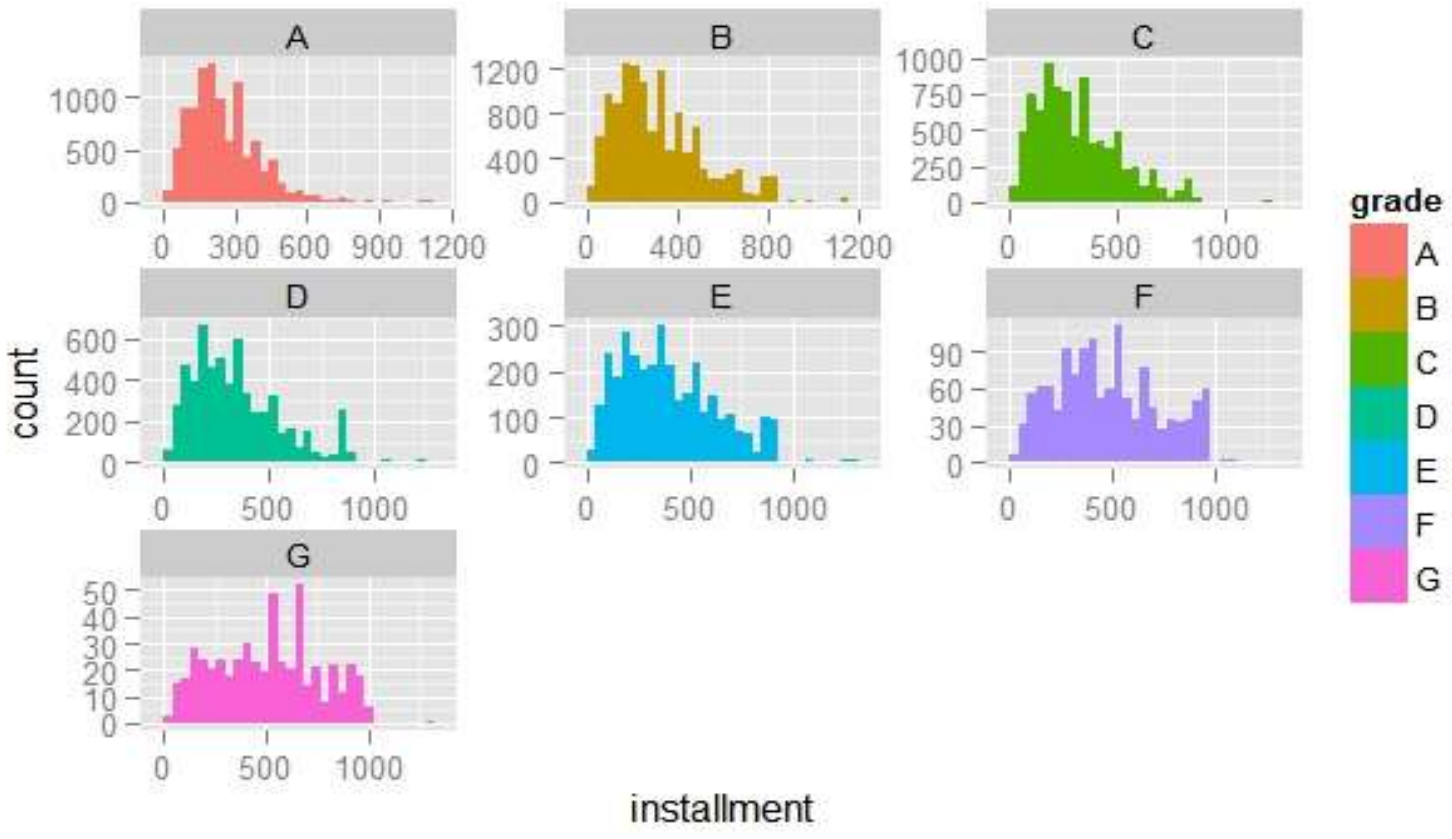
输出结果


A
B
C
D
E
F
G
1500
1000
500
0
count
0
500
1000
grade
A
B
C
D
E
F
G
installment

自动调整

- `p+geom_histogram()+facet_wrap(~grade,scales="free")`

Scale=”free”可以自动调整坐标刻度：包括free_y与free_x



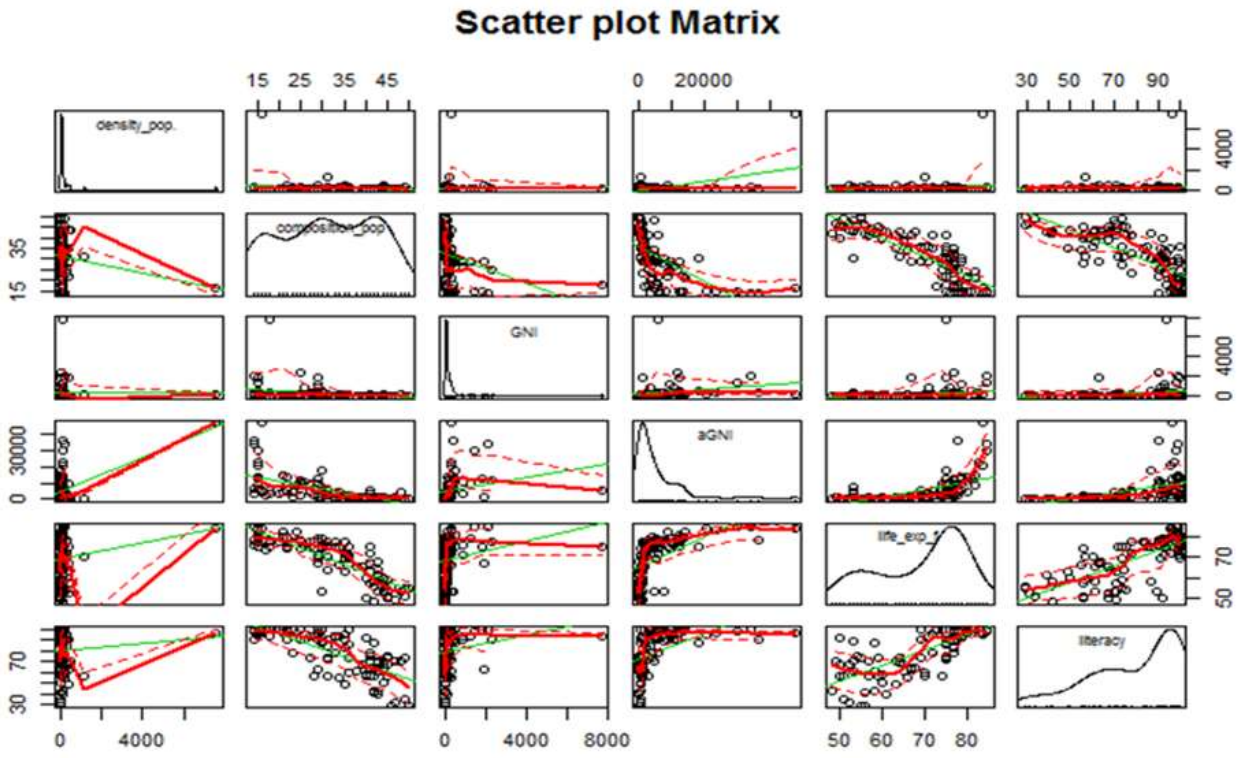
3.4 高维数据的展示

高维数据的展示

- 高维数据展示的第一步，通常是采用种种技术手段将n维空间中的点降至三维以下，从而能够绘制于平面设备上。
- 散点图矩阵
- 相关系数矩阵
- 聚类热图
- 平行坐标图（轮廓图）
- 雷达图
- 脸谱图

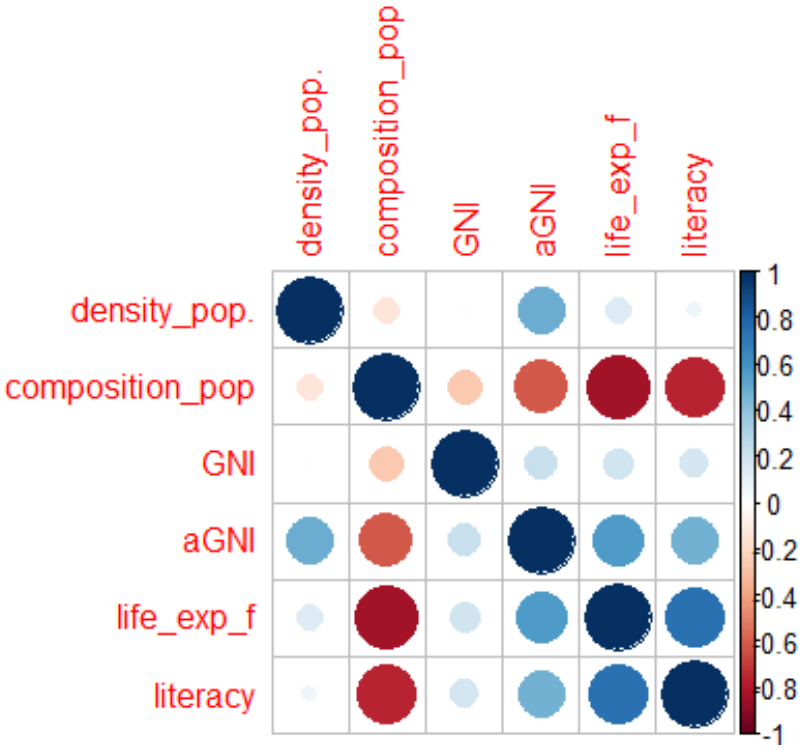
散点图矩阵

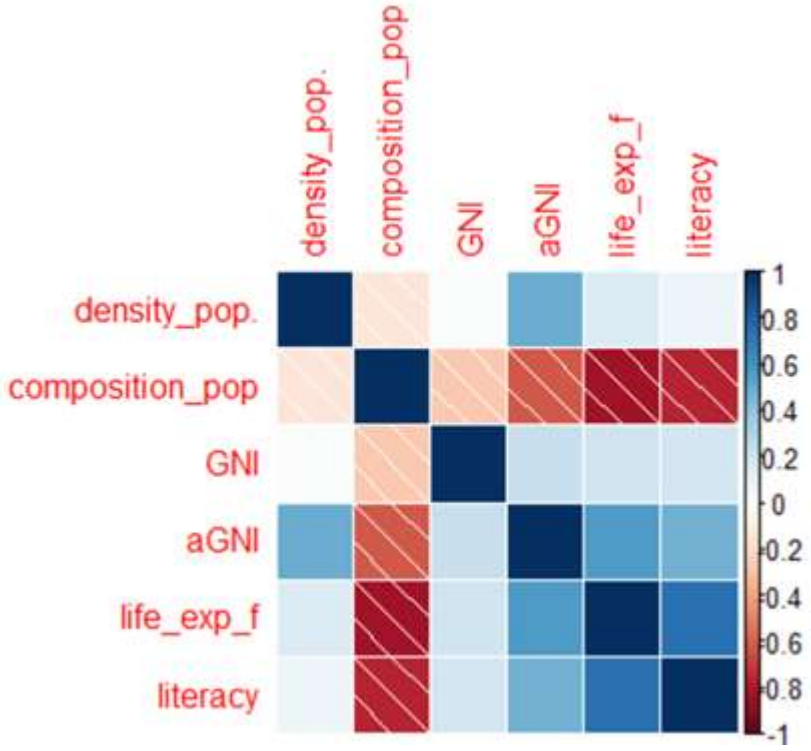
- 散点图用点在坐标系上的位置展示多个数值变量的两两分布
- 适用数据：多个定量数据
- 主要功能：探索多个变量间两两的相关关系
- R的car包中的scatterplotMatrix函数可以绘制散点图矩阵
  - #scatterplotm
  - Library(car)
  - scatterplotMa



相关系数矩阵

- corrplot包中的corrplot函数提供了形式多样的可视化相关矩阵的方法

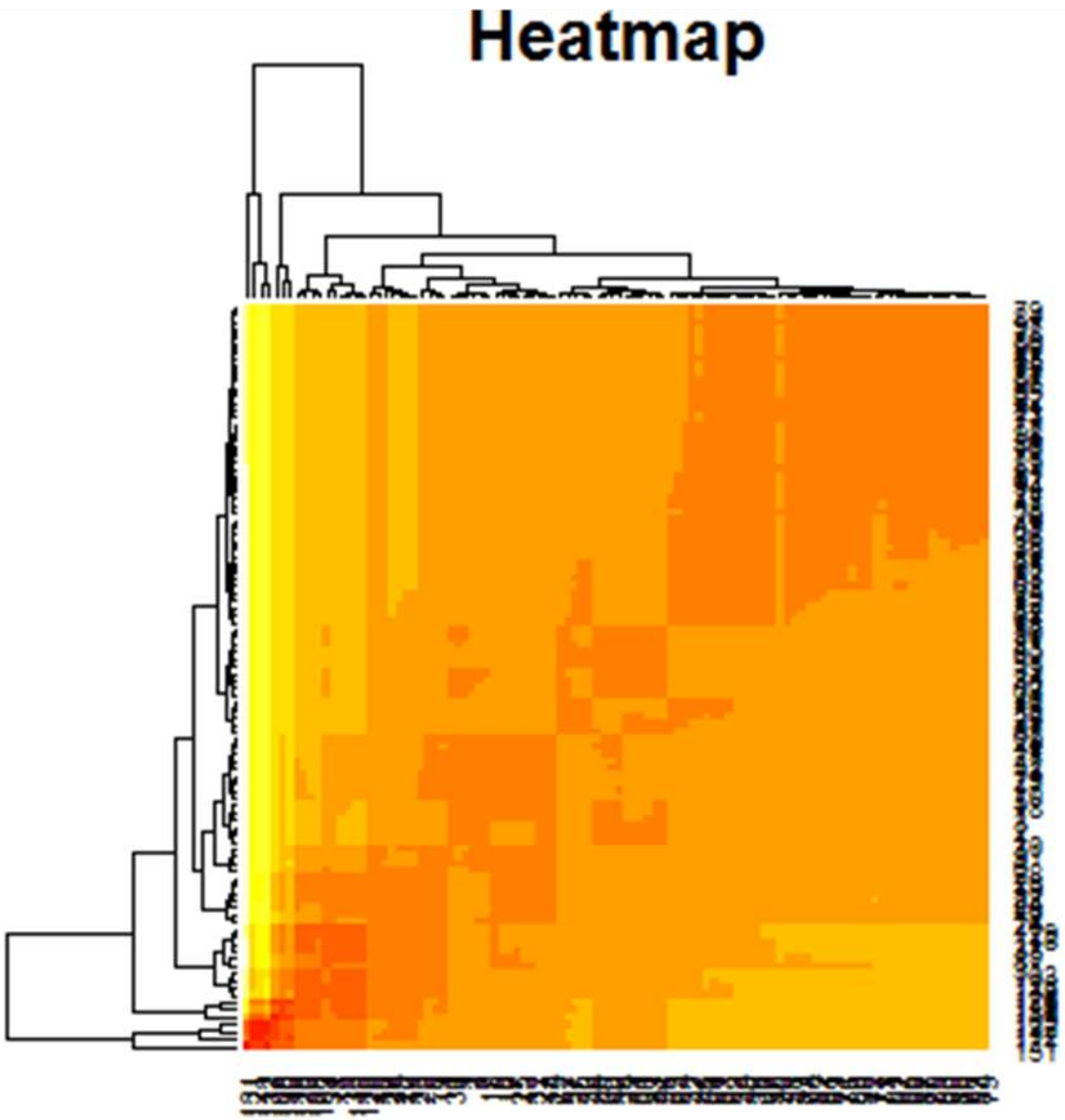




聚类热图

- 在聚类分析技术中，通常用热图来大致区分出类的数量以及哪些观测点同属一类。热图的绘制方式与相关系数矩阵图及其类似，都是通过将某个数值映射到矩阵中的连续变化的颜色上。不同的是，热图使用的数值是某两个观测点间的欧式距离，而相关系数图使用的是两个变量的相关系数。前者站在观测值的角度，后者站在变量的角度。
- 以worldbank中的keyindicators1为例，依据density_pop, aGNI, life_exp_f, literacy四个变量计算欧式距离，使用heatmap函数绘制。

热图

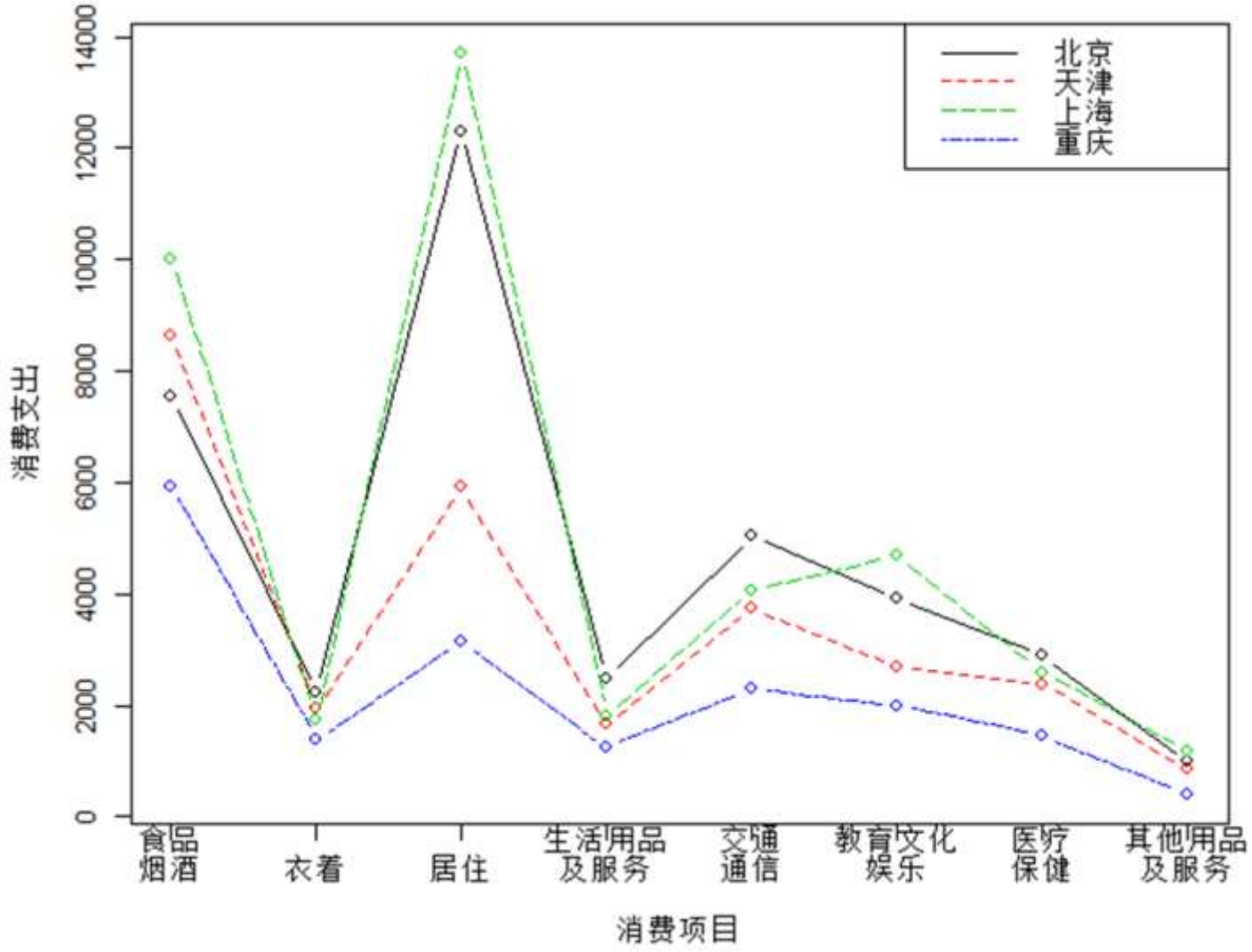


热图中区块的颜色深浅表示两个观测点的距离远近，邻近的点对应的方格的颜色更深，而远处的点对应的方格颜色浅。

平行坐标图（轮廓图）

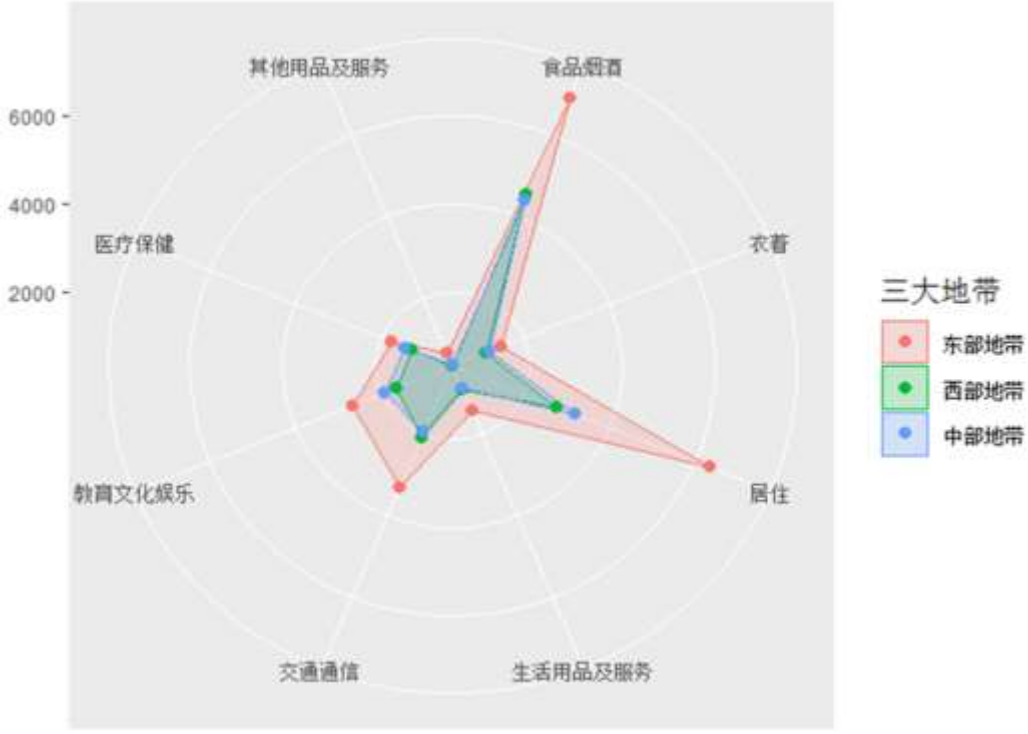
- 平行坐标图将每个观测值在不同变量上的取值连接成线，曲线走势相似的观测点倾向于为同一类
- 平行坐标轴的绘制原理是在横轴上选择几个等距的点表示不同的变量，变量的取值（或者是经过标准化处理的值）被映射到纵坐标上。这样，一个n维随机变量的一个观测就可以表示为n个点，将每个观测的点依次连接起来就形成了平行坐标图。
- 适用数据：多个定量变量
- 主要功能：可用于研究多个样本在多个变量上的相似程度或变量间的相互关系
- 用matplot函数绘制4个直辖市的8项消费支出的轮廓图

平行坐标图（轮廓图）


北京
天津
上海
重庆
消费支出
0
2000
4000
6000
8000
10000
12000
14000
食品烟酒
衣着
居住
生活用品及服务
交通通信
教育文化娱乐
医疗保健
其他用品及服务
消费项目

雷达图

- 假设有P个变量，我们可以从一个点出发，每个变量用一条射线表示，P个变量形成P条射线，每个样本点在P个变量上的取值连接成线，即围成一个区域，多个样本围成多个区域，就是雷达图。
- P个变量的计量单位可能不同，数值的量级往往差异很大，每条坐标轴的刻度需要根据每个变量单独确定，因此，不同坐标轴的刻度是不可比的。
- 利用雷达图可以研究多个样
- 利用fmsb包中的redarchart函数绘制雷达图

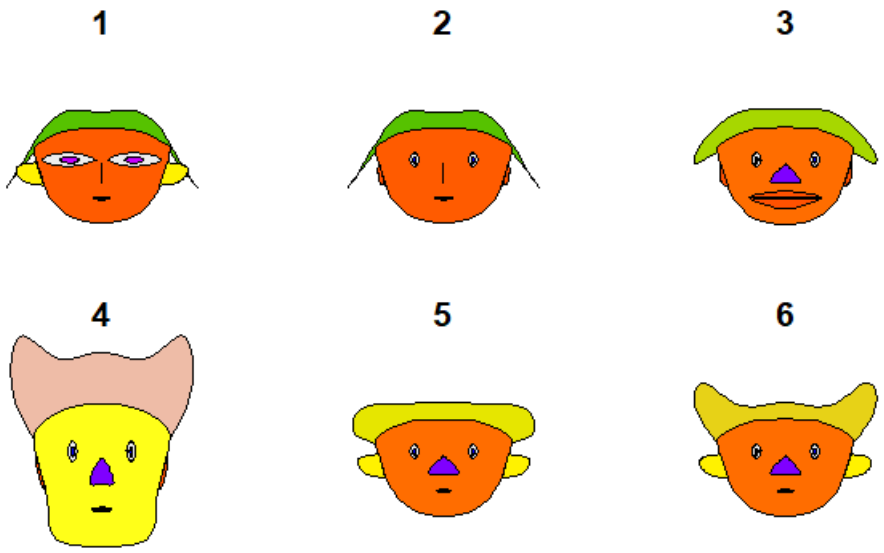


脸谱图

- 脸谱图由美国科学家Chernoff首先提出。脸谱图将P个变量用人脸部位的形状或大小来表征，通过对脸谱的分析，可根据P个变量对样本进行归类或比较研究。
- 按照Chernoff提出的画法，由15个变量描绘脸部的特征，若实际变量更多，多出的将被忽略；若实际变量较少，变量将被重复使用，这时某个变量可能同时描述样本的几个特征。
- 绘制脸谱图的R包有aplpack，symbols，DescTools，TeachingDemos等

变量	面部特征	变量	面部特征	变量	面部特征
1	脸的高度	6	笑容曲线	11	发型
2	脸的宽度	7	眼睛高度	12	鼻子高度
3	脸的形状	8	眼睛宽度	13	鼻子宽度
4	嘴的高度	9	头发高度	14	耳朵宽度
5	嘴的宽度	10	头发宽度	15	耳朵高度

脸谱图：特殊群体分析：90岁以上老人

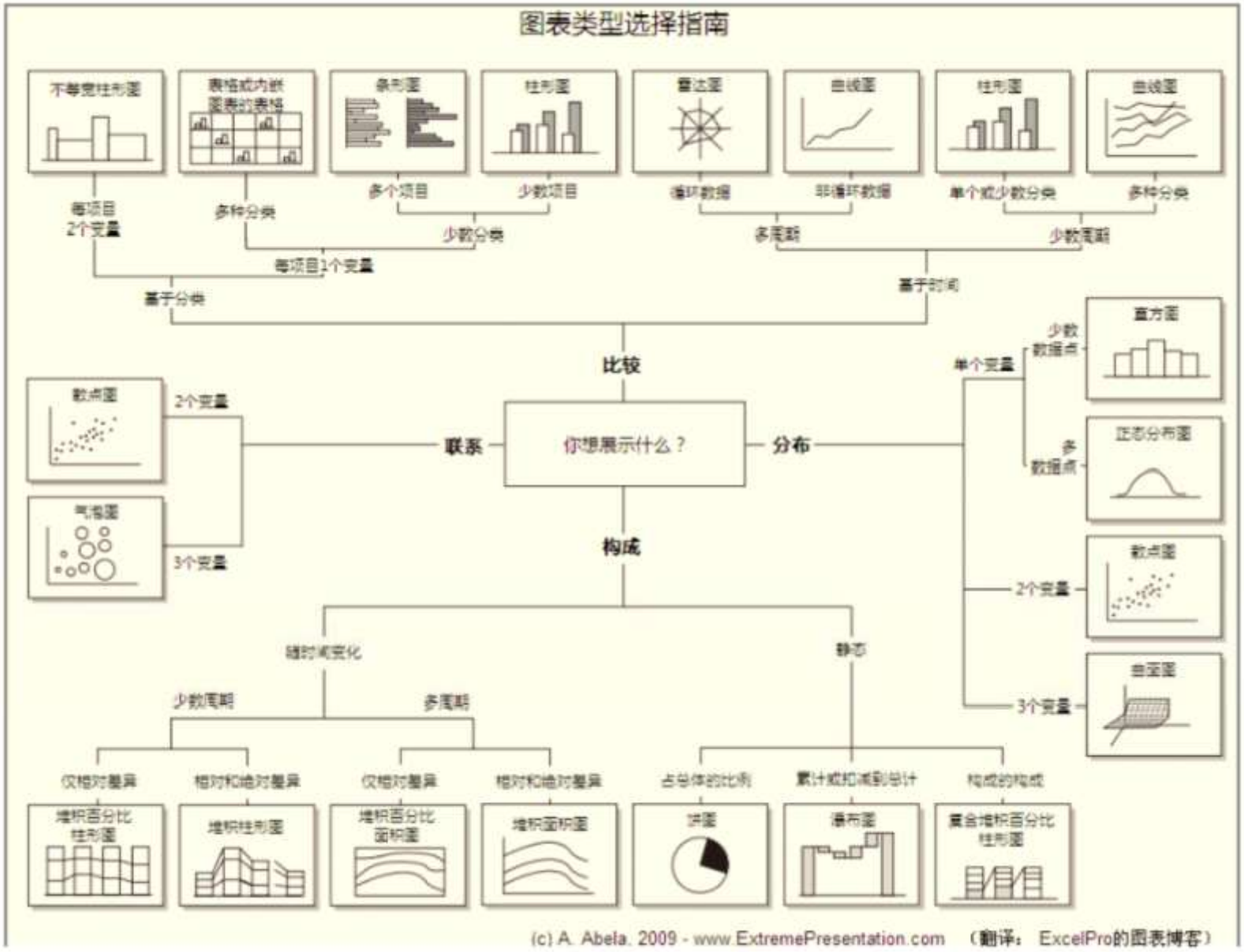


90岁以上的有6位老人：

- 几位老人都不笑，记忆力都为0
- 1号老人眼睛大，他有点抑郁
- 3号老人资产最多（嘴的宽度）
- 4号老人头收入最多（头发高度）
- 4号和6号老人发型时髦，他们参与娱乐社交多

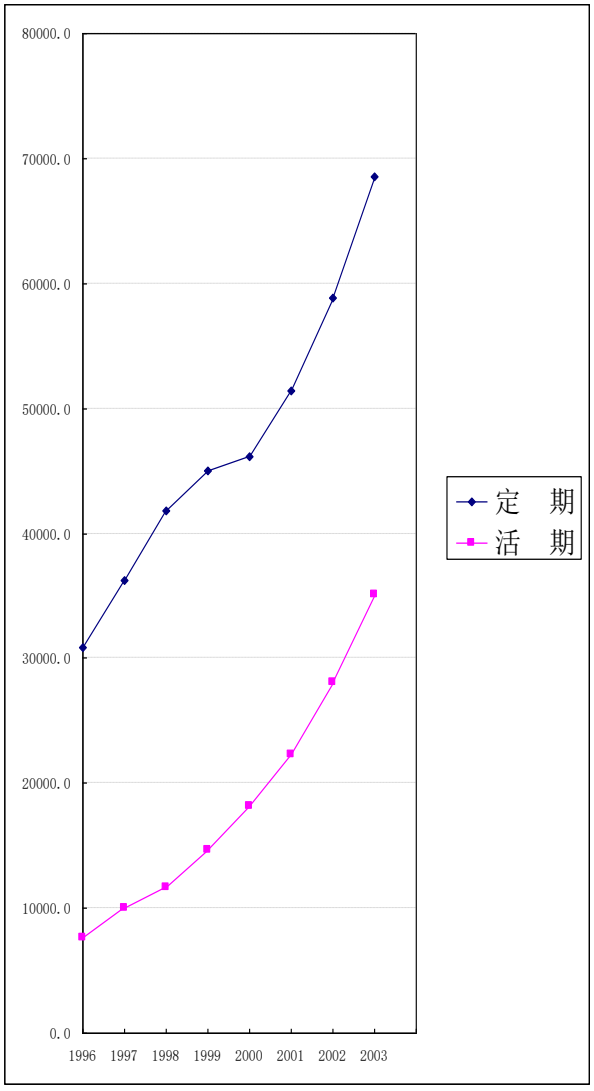
特征	变量
Height of face	Pension
Width of face	Salary
Structure of face	Zy_ic
Height of mouth	Money_child
Width of mouth	Property
Smiling	Memory
Height of eyes	Status
Width of eyes	Depress
Height of hair	Inc_total
Width of hair	Freq_vol
Style of hair	Freq_ent
Height of nose	Age_d
Width of nose	Social_kinds.y
Width of ear	Family_num
Height of ear	Cognition

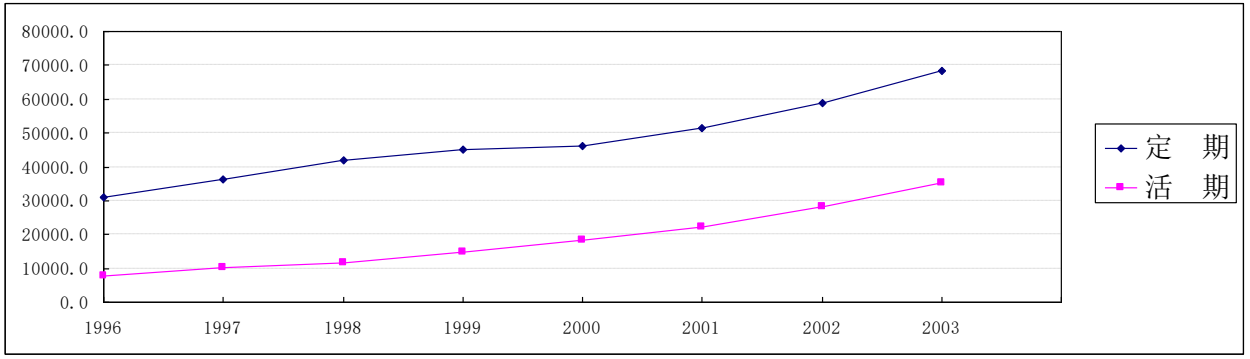
图表类型选择指南

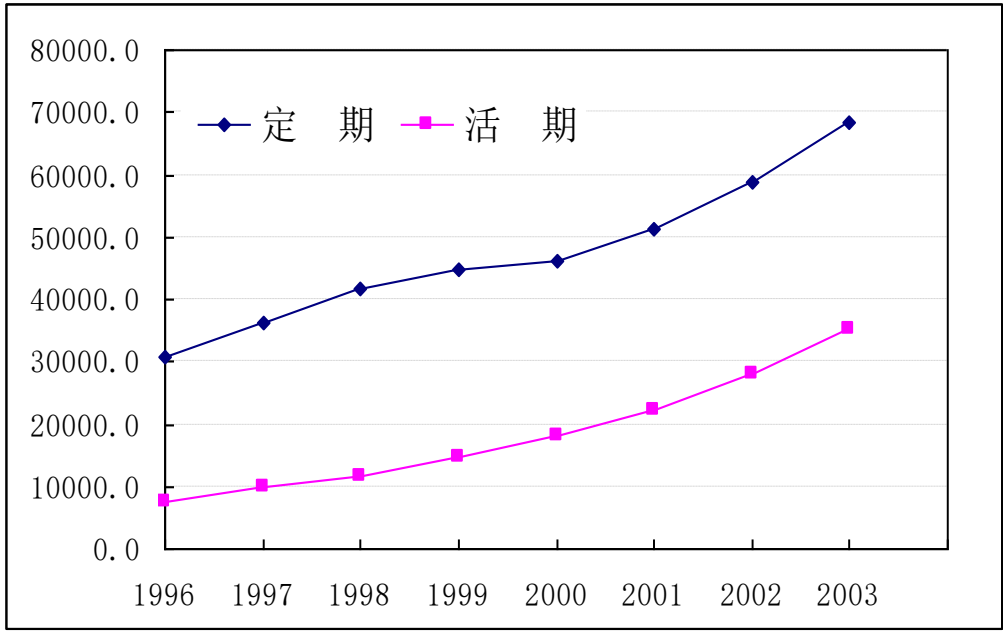

图表类型选择指南
不等宽柱形图
表格或内嵌图表的表格
条形图
柱形图
雷达图
曲线图
柱形图
曲线图
每项目2个变量
多种分类
多个项目
少数项目
循环数据
非循环数据
单个或少数分类
多种分类
少数分类
多周期
少数周期
每项目1个变量
基于分类
基于时间
比较
你想展示什么？
联系
分布
构成
散点图
2个变量
气泡图
3个变量
单个变量
少数数据点
多数据点
直方图
正态分布图
2个变量
散点图
3个变量
曲面图
随时间变化
静态
少数周期
多周期
仅相对差异
相对和绝对差异
仅相对差异
相对和绝对差异
占总体的比例
累计或扣减到总计
构成的构成
堆积百分比柱形图
堆积柱形图
堆积百分比面积图
堆积面积图
饼图
瀑布图
复合堆积百分比柱形图
(c) A. Abela. 2009 - www.ExtremePresentation.com （翻译： ExcelPro的图表博客）

合理使用图表

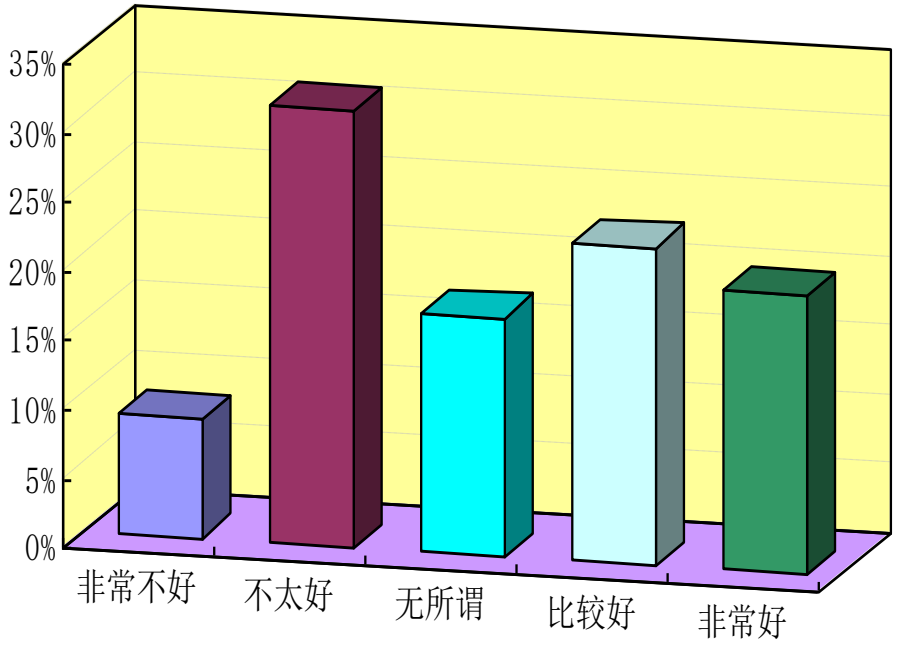
注意图形比例

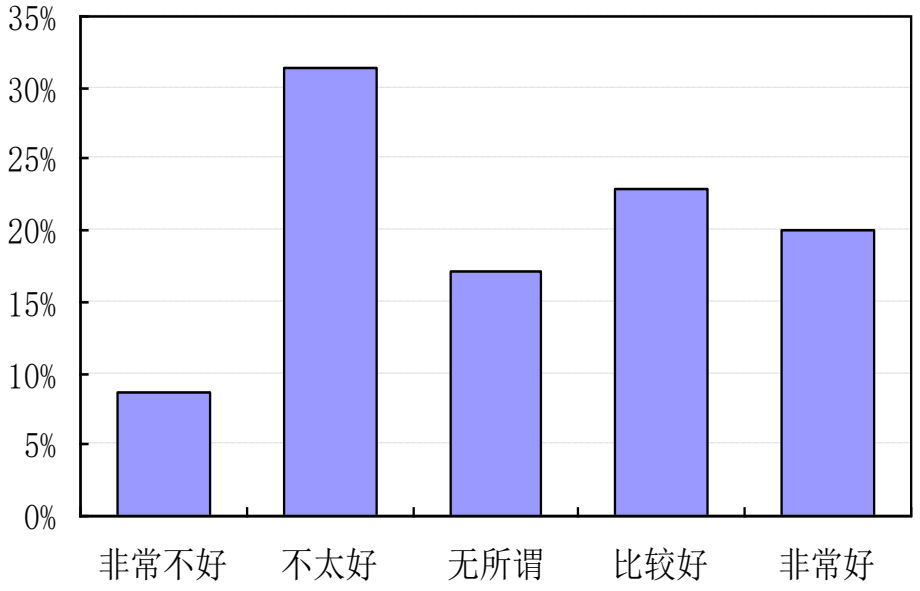

80000.0
70000.0
60000.0
50000.0
40000.0
30000.0
20000.0
10000.0
0.0
1996
1997
1998
1999
2000
2001
2002
2003
定　期
活　期


80000.0
70000.0
60000.0
50000.0
40000.0
30000.0
20000.0
10000.0
0.0
1996
1997
1998
1999
2000
2001
2002
2003
定　期
活　期


80000.0
70000.0
60000.0
50000.0
40000.0
30000.0
20000.0
10000.0
0.0
1996
1997
1998
1999
2000
2001
2002
2003
定　期
活　期

慎用三维效果

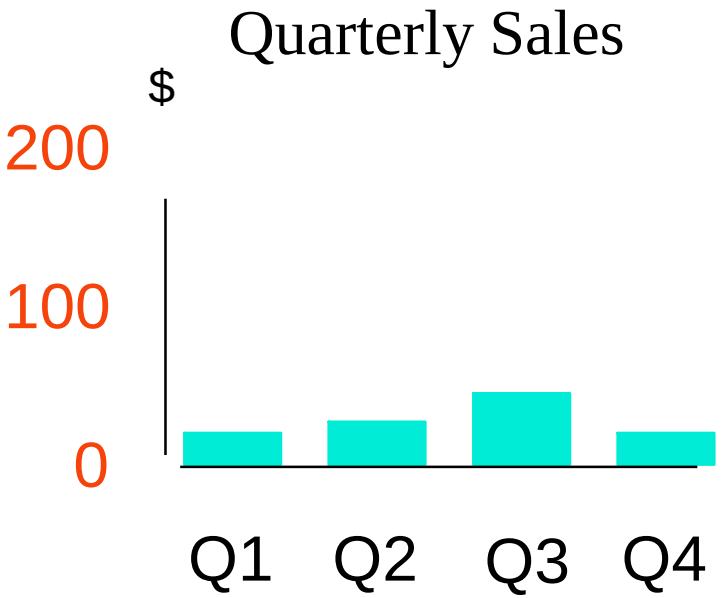




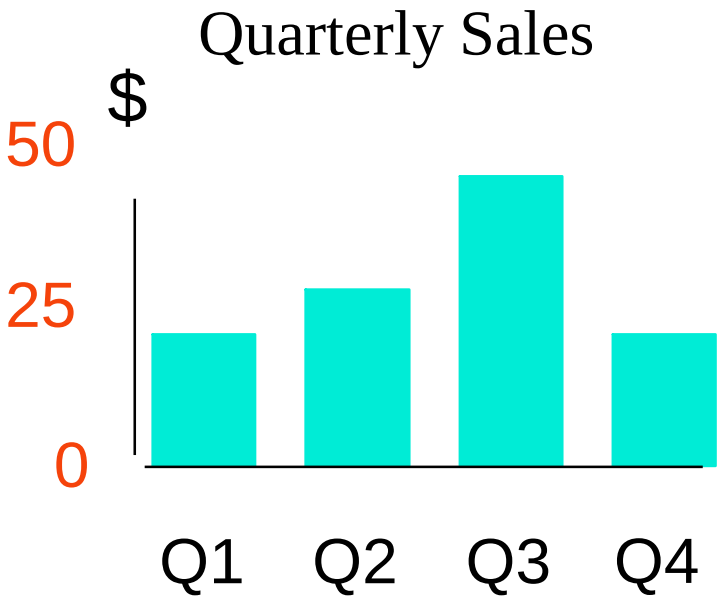
不必要的三维效果：三维图形可能比二维图形更能吸引读者的注意，但只能用来反映变化的趋势，不能用来进行精确的比较。

选择恰当的数轴刻度

不好的图形

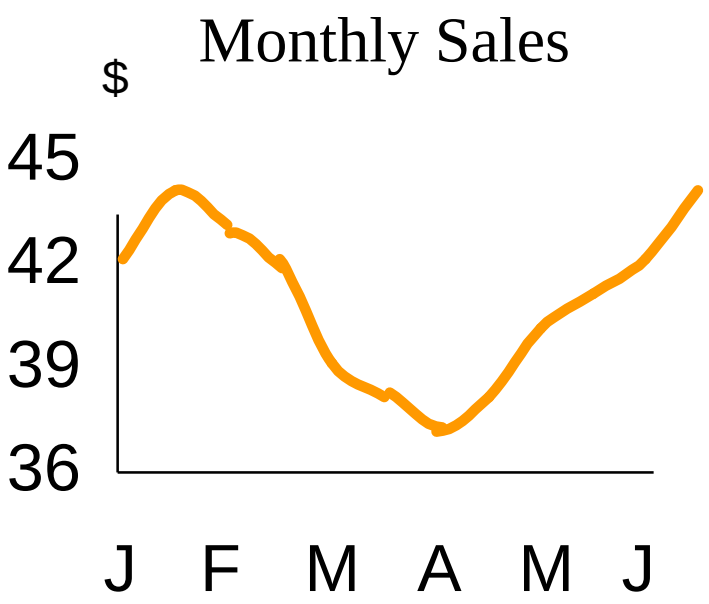


好的图形

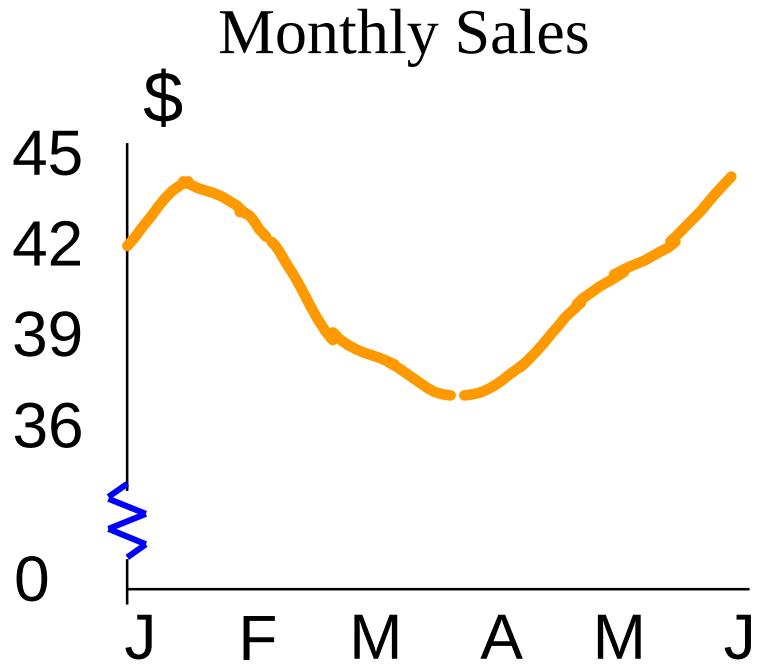


选择恰当的数轴刻度

不好的图形



好的图形



找问题


2000–2008年储蓄率构成
60%
50%
40%
30%
20%
10%
0%
16.5%
16.2%
18.6%
18.1%
19.3%
21.2%
21.8%
21.8%
17.3%
2000年
2001年
2002年
2003年
2004年
2005年
2006年
2007年
2008年
企业储蓄率
政府储蓄率
居民储蓄率

鉴别图表优劣的准则

- **爱德华·R·塔夫特（Edward R. Tufte）在其著作The Visual Display of Quantitative Information（1983）中用“图优性”（graphical excellency）来描述一张好图。**
- **一张好的图表应包括以下基本特征**
  - 显示数据
  - 让读者把注意力集中在图表的内容上，而不是制作图表的程序上
  - 避免歪曲
  - 强调数据之间的比较
  - 服务于一个明确的目的
  - 有对图表的统计描述和文字说明

总结与展望

统计制图总结

- 分析数据，选用合适的图
- 必要时，连续变离散，定量变定性
- 尽量让图形简单并美观大方
- 指标值图形化
- 优秀的可视化分析报告：搭建合理的分析框架，层层深入

数据分析报告的逻辑性

- 基本模式：
- 提出问题——吸引眼球
- 问题所对应的现象的展示——冲击性强
- 现象的多样化及影响——丰富层次
- 原因的剖析——解答疑惑
- 结论和建议——解决问题
- ……

课程结束，谢谢大家！

祝愿大家在市调赛取得优秀成绩！