北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Data Analytics with R
语言数据分析

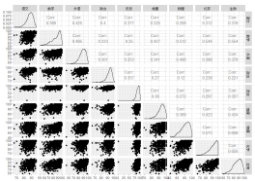
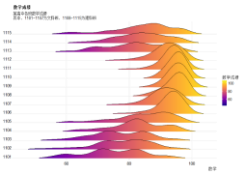
观数以形

艾新波 / 2018 • 北京

课程体系

- R语言数据分析
  - 上部：论道 
    - 第1章 气象万千、数以等观
    - 第2章 所谓学习、归类而已
    - 第3章 格言联璧话学习
    - 第4章 源于数学、归于工程
  - 中部：执具 
    - 第5章 工欲善其事必先利其器
    - 第6章 基础编程
    - 第7章 数据对象
  -
    - 第8章 人人都爱tidyverse
    - 第9章 最美不过数据框
  - 下部 博术 
    - 第10章 观数以形
    - 第11章 相随相伴、谓之关联
    - 第12章 既是世间法、自当有分别
    - 第13章 方以类聚、物以群分
    - 第14章 庐山烟雨浙江潮

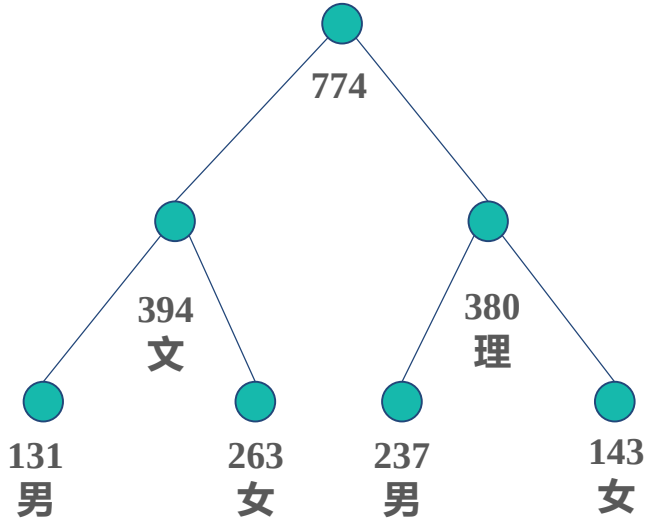
二维数据空间的形态

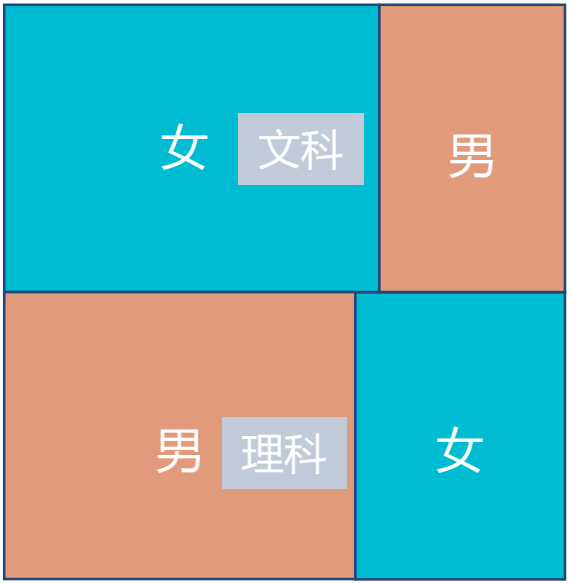
类别	刻画方法	图形示例
类别变量vs类别变量 (因子vs因子)	矩形树图	
数值变量vs数值变量 (向量vs向量)	散点图、 相关系数图	
类别变量vs数值变量 (因子vs向量)	分组绘制箱线图/ 直方图/概率密度图等	

离散变量vs离散变量：矩形树图

矩形树图采用嵌套的矩形来表达层级数据

数值大小映射为面积和颜色

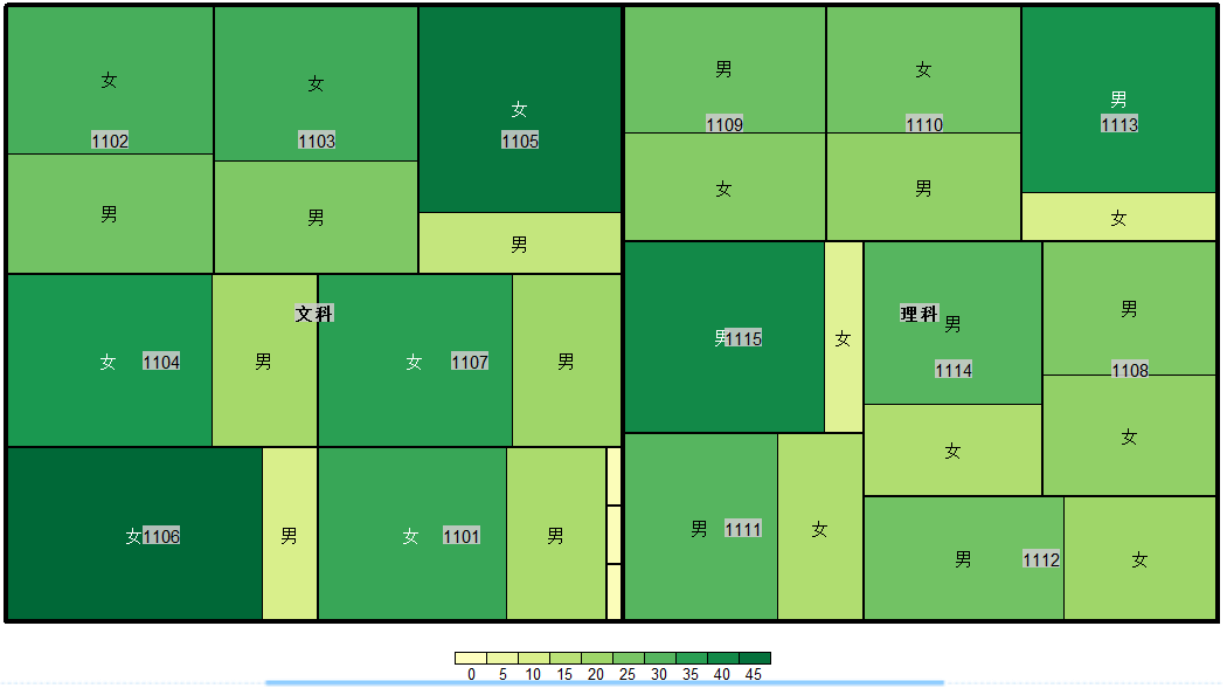




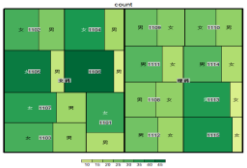
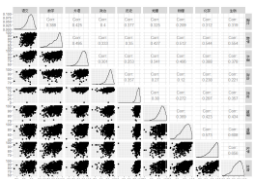
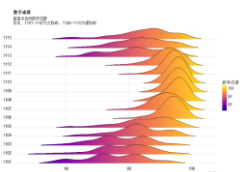
离散变量vs离散变量：矩形树图

```
library(treemap)
cjb %>%
  group_by(wlfk, bj, xb) %>%
  summarise(count = n()) %>%
  as.data.frame() %>%
  treemap(
    index = c("wlfk", "bj", "xb"),
    vSize = "count",
    vColor = "count",
    type = "value"
  )
```

离散变量vs离散变量：矩形树图



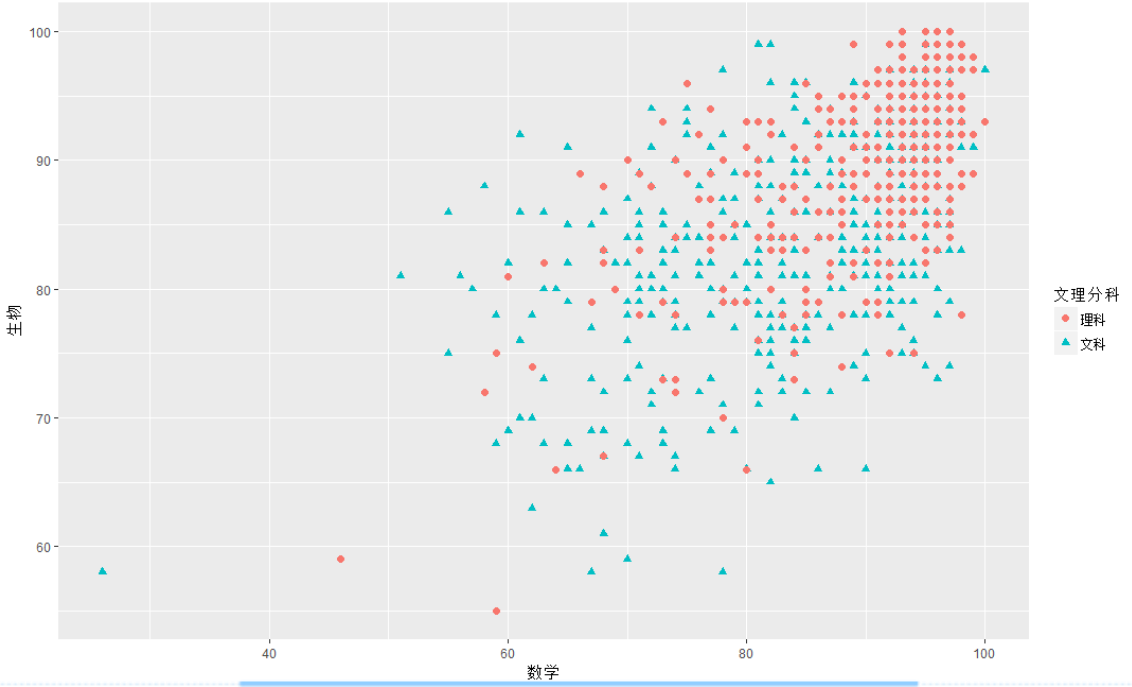
二维数据空间的形态

类别	刻画方法	图形示例	
类别变量vs类别变量 (因子vs因子)	矩形树图		✓
数值变量vs数值变量 (向量vs向量)	散点图、 相关系数图		○
类别变量vs数值变量 (因子vs向量)	分组绘制箱线图/ 直方图/概率密度图等		○

数值变量 vs 数值变量：散点图

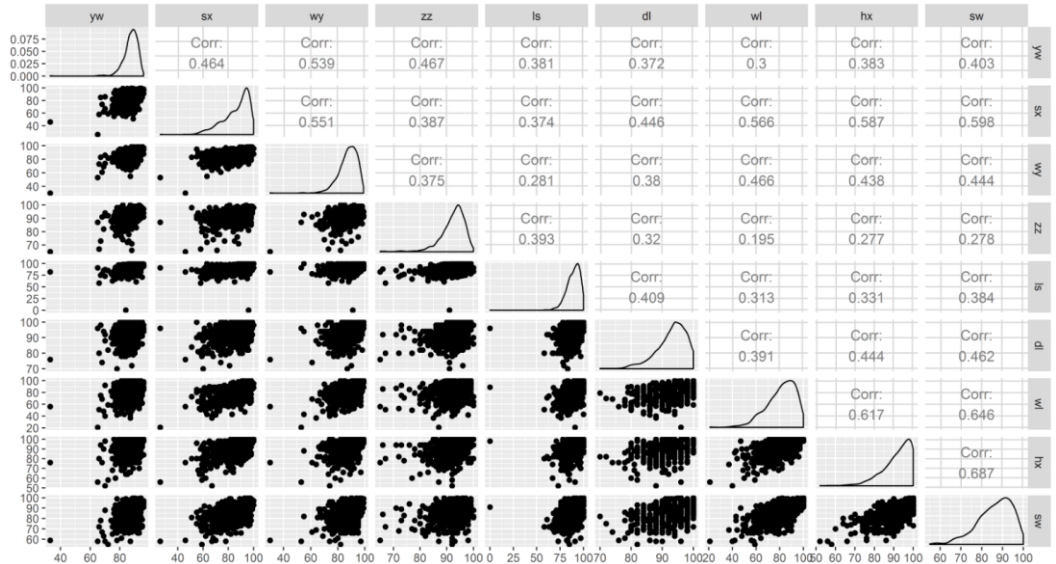
```
library(ggplot2)
ggplot(cjb,
       aes(x = sx,
           y = sw,
           shape = wlfk,
           colour = wlfk)) +
  geom_point(size = 2) +
  labs(x = "数学",
       y = "生物",
       colour = "文理分科",
       shape = "文理分科")
```

数值变量 vs 数值变量：散点图



数值变量 vs 数值变量：散点图矩阵

```
GGally::ggpairs(cjb, columns = 4:12)
```



数值变量vs数值变量：相关系数

共变与相关：

当一个变量增大，另一个变量随之增大（或减小），称这种现象为共变，或相关（correlation）

两个变量之间具有共变现象，称为具有相关关系

在没有特别说明的情况下，相关系数一般是指线性相关系数

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

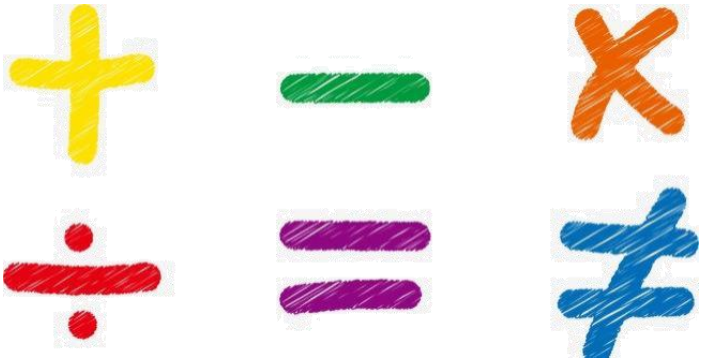
容易看到，相关系数的分子即为协方差(1/(n-1))

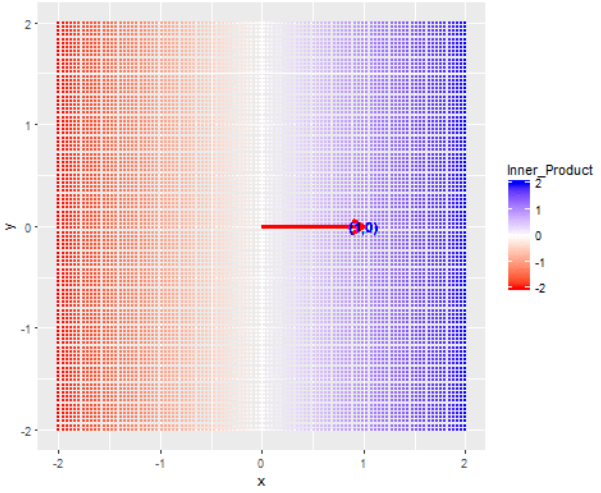
顾名思义，协方差之“协”，在于刻画其变化的相似性

数值变量vs数值变量：相关系数

结合内积的概念，考虑协方差之“协”

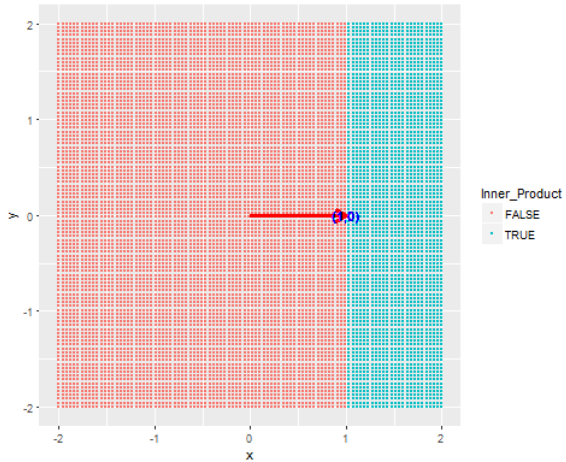
$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$



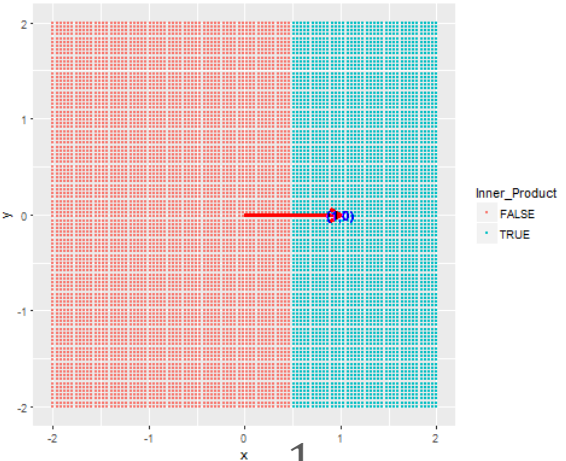


数值变量vs数值变量：相关系数

进一步延伸$w^T x + b = 0$，为下一步的分类超平面做准备



$b = -1$



$b = -\frac{1}{2}$

相关系数的R语言实现

```
(cor_coef <- cor(cjb[, 4:12]))
#>      yw   sx   wy   zz   ls   dl   wl   hx   sw
#> yw 1.00 0.46 0.54 0.47 0.38 0.37 0.30 0.38 0.40
#> sx 0.46 1.00 0.55 0.39 0.37 0.45 0.57 0.59 0.60
#> wy 0.54 0.55 1.00 0.37 0.28 0.38 0.47 0.44 0.44
#> zz 0.47 0.39 0.37 1.00 0.39 0.32 0.20 0.28 0.28
#> ls 0.38 0.37 0.28 0.39 1.00 0.41 0.31 0.33 0.38
#> dl 0.37 0.45 0.38 0.32 0.41 1.00 0.39 0.44 0.46
#> wl 0.30 0.57 0.47 0.20 0.31 0.39 1.00 0.62 0.65
#> hx 0.38 0.59 0.44 0.28 0.33 0.44 0.62 1.00 0.69
#> sw 0.40 0.60 0.44 0.28 0.38 0.46 0.65 0.69 1.00
```

相关系数的R语言实现

```
(cor_coef <- cor(cjb[, 4:12]))
cor_coef %>%
  as.data.frame() %>%
  rownames_to_column(var = "km1") %>%
  gather(key = km2, value = cor_num, -km1) %>%
  mutate(cor_level = cut(cor_num,
                         breaks= c(0, 0.3, 0.5, 0.8, 1),
                         right = FALSE)) %>%
  ggplot(aes(x = km1, y = km2, fill = cor_level)) +
  geom_tile(colour="white", size = 1.5) +
  geom_text(aes(label = format(cor_num, digits = 2))) +
  scale_fill_brewer(palette = "YlGn",name="相关系数区间")
```

相关系数的R语言实现

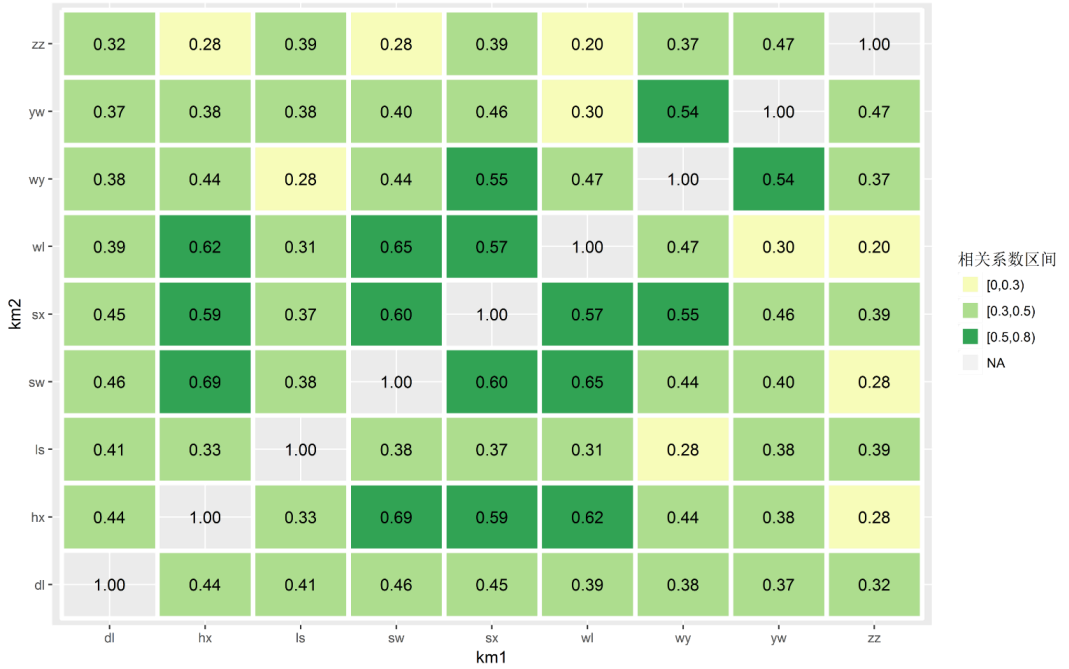
	km1	yw	sx	wy	zz	ls	dl	wl	hx	sw
1	yw	1.00	0.46	0.54	0.47	0.38	0.37	0.30	0.38	0.40
2	sx	0.46	1.00	0.55	0.39	0.37	0.45	0.57	0.59	0.60
3	wy	0.54	0.55	1.00	0.37	0.28	0.38	0.47	0.44	0.44
4	zz	0.47	0.39	0.37	1.00	0.39	0.32	0.20	0.28	0.28
5	ls	0.38	0.37	0.28	0.39	1.00	0.41	0.31	0.33	0.38
6	dl	0.37	0.45	0.38	0.32	0.41	1.00	0.39	0.44	0.46
7	wl	0.30	0.57	0.47	0.20	0.31	0.39	1.00	0.62	0.65
8	hx	0.38	0.59	0.44	0.28	0.33	0.44	0.62	1.00	0.69
9	sw	0.40	0.60	0.44	0.28	0.38	0.46	0.65	0.69	1.00

	km1	km2	cor_num	cor_level
1	yw	yw	1.00	<NA>
2	sx	yw	0.46	[0.3,0.5)
3	wy	yw	0.54	[0.5,0.8)
4	zz	yw	0.47	[0.3,0.5)
5	ls	yw	0.38	[0.3,0.5)
6	dl	yw	0.37	[0.3,0.5)
7	wl	yw	0.30	[0.3,0.5)
8	hx	yw	0.38	[0.3,0.5)
9	sw	yw	0.40	[0.3,0.5)
10	yw	sx	0.46	[0.3,0.5)
11	sx	sx	1.00	<NA>
12	wy	sx	0.55	[0.5,0.8)
13	zz	sx	0.39	[0.3,0.5)
14	ls	sx	0.37	[0.3,0.5)
15	dl	sx	0.45	[0.3,0.5)
16	wl	sx	0.57	[0.5,0.8)
17	hx	sx	0.59	[0.5,0.8)
18	sw	sx	0.60	[0.5,0.8)
19	yw	wy	0.54	[0.5,0.8)
20	sx	wy	0.55	[0.5,0.8)

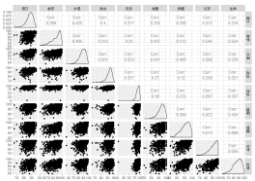
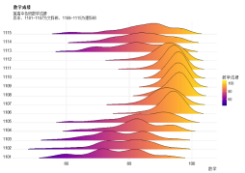
相关系数的R语言实现

```
(cor_coef <- cor(cjb[, 4:12]))
cor_coef %>%
  as.data.frame() %>%
  rownames_to_column(var = "km1") %>%
  gather(key = km2, value = cor_num, -km1) %>%
  mutate(cor_level = cut(cor_num,
                         breaks= c(0, 0.3, 0.5, 0.8, 1),
                         right = FALSE)) %>%
  ggplot(aes(x = km1, y = km2, fill = cor_level)) +
  geom_tile(colour="white", size = 1.5) +
  geom_text(aes(label = format(cor_num, digits = 2))) +
  scale_fill_brewer(palette = "YlGn",name="相关系数区间")
```

相关系数的R语言实现


zz 0.32 0.28 0.39 0.28 0.39 0.20 0.37 0.47 1.00
yw 0.37 0.38 0.38 0.40 0.46 0.30 0.54 1.00 0.47
wy 0.38 0.44 0.28 0.44 0.55 0.47 1.00 0.54 0.37
wl 0.39 0.62 0.31 0.65 0.57 1.00 0.47 0.30 0.20
sx 0.45 0.59 0.37 0.60 1.00 0.57 0.55 0.46 0.39
sw 0.46 0.69 0.38 1.00 0.60 0.65 0.44 0.40 0.28
ls 0.41 0.33 1.00 0.38 0.37 0.31 0.28 0.38 0.39
hx 0.44 1.00 0.33 0.69 0.59 0.62 0.44 0.38 0.28
dl 1.00 0.44 0.41 0.46 0.45 0.39 0.38 0.37 0.32
km2
dl hx ls sw sx wl wy yw zz
km1
相关系数区间
[0,0.3)
[0.3,0.5)
[0.5,0.8)
NA

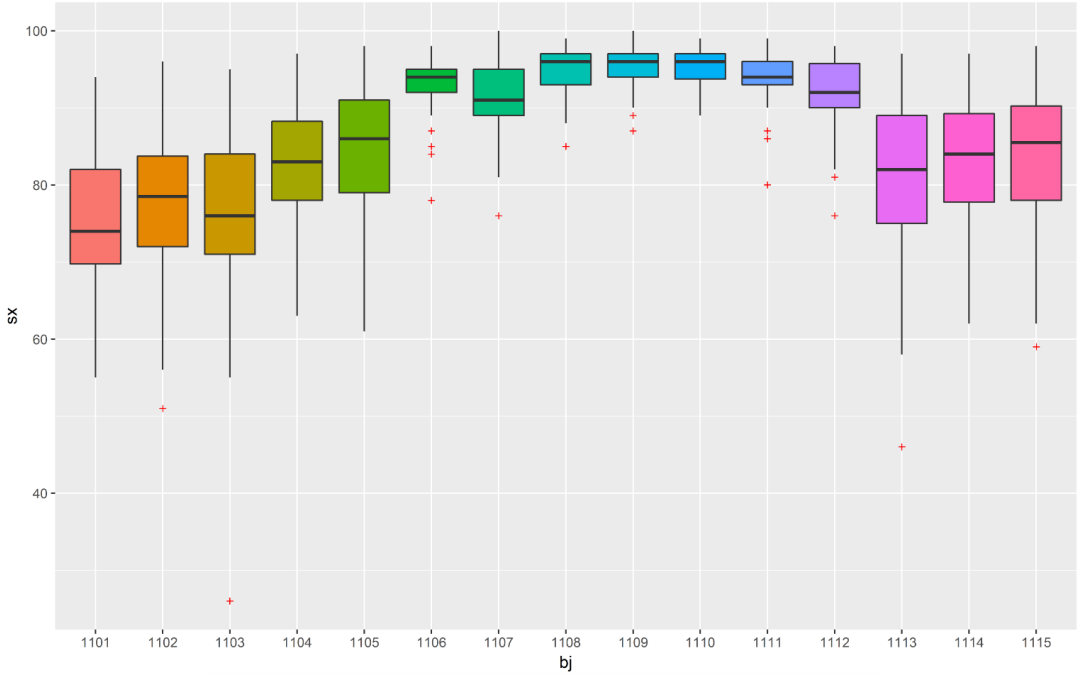
二维数据空间的形态

类别	刻画方法	图形示例	
类别变量vs类别变量 (因子vs因子)	矩形树图		✓
数值变量vs数值变量 (向量vs向量)	散点图、 相关系数图		✓
类别变量vs数值变量 (因子vs向量)	分组绘制箱线图/ 直方图/概率密度图等		○

类别变量vs数值变量：分组绘图

```
#分组绘制箱线图
#看看不同班级数学成绩的分布
library(ggplot2)
ggplot(cjb, aes(x = bj,
                y = sx,
                fill = bj)) +
  geom_boxplot(outlier.colour = "red",
               outlier.shape = 3,
               outlier.size = 1) +
  labs(x = "班级", y = "数学成绩") +
  theme(legend.position = "none")
```

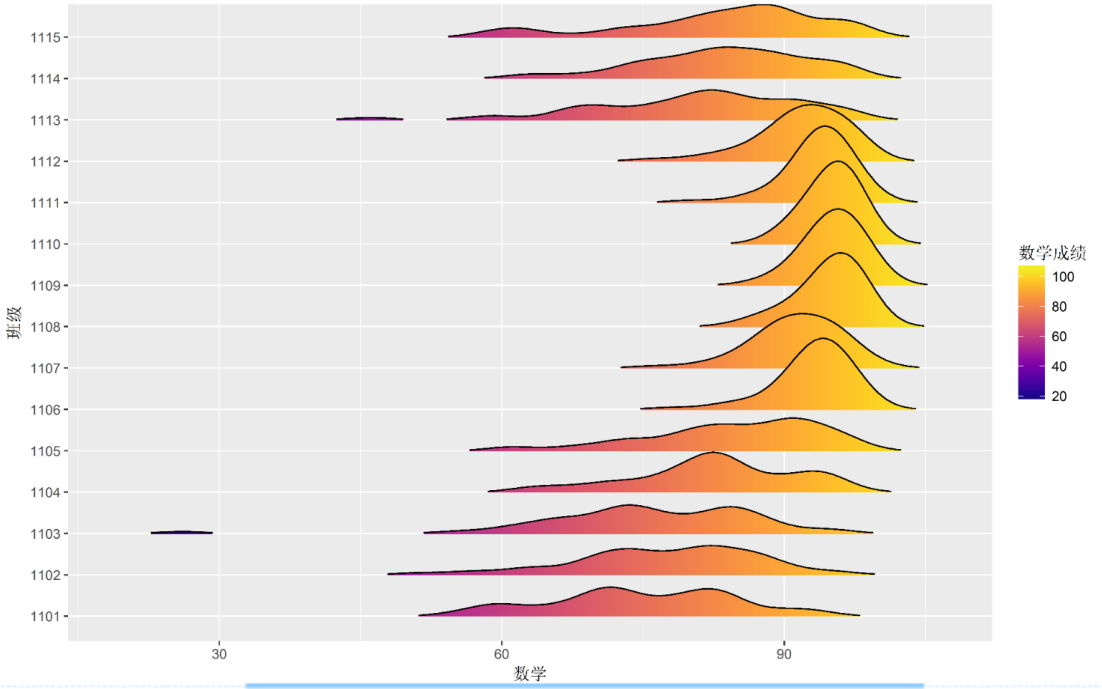
类别变量vs数值变量：分组绘图



类别变量vs数值变量：分组绘图

```
library(ggridges)#绘制层峦叠嶂图
library(viridis)#采用其中的颜色
ggplot(cjb, aes(x = sx, y = bj, fill = ..x..)) +
  geom_density_ridges_gradient(
    scale = 2,
    rel_min_height = 0.01,
    gradient_lwd = 1) +
  scale_fill_viridis(
    name = "数学成绩",
    option = "C") +
  labs(x = "数学", y = "班级")
```

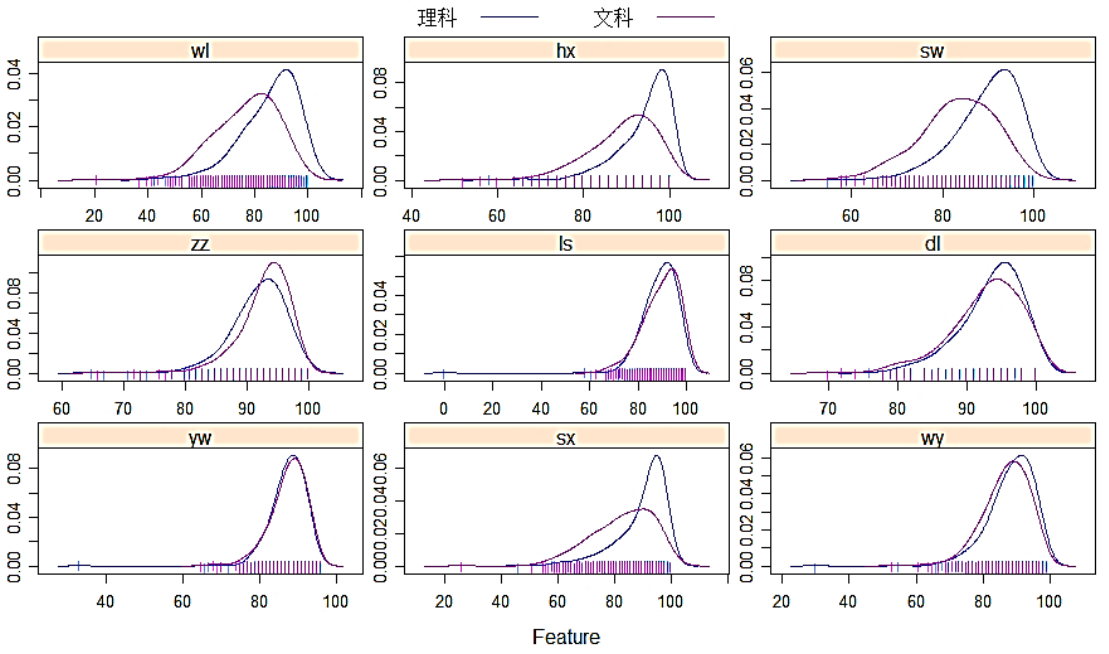
类别变量vs数值变量：分组绘图



类别变量vs数值变量：分组绘图

```
#对于分类问题而言，在进行数据描述时
#最关键的，当属因变量vs自变量了
library(caret)
featurePlot(
  x = cjb[, 4:12],
  y = cjb$wlfk,
  plot = "density",
  scales = list(
    x = list(relation = "free"),
    y = list(relation = "free")),
  adjust = 1.5,
  pch = "|",
  auto.key = list(columns = 2))
```

类别变量vs数值变量：分组绘图



延伸阅读： CHEAT SHEET(ggplot2)

ONE VARIABLE continuous

c <- ggplot(mpg, aes(hwy)); c2 <- ggplot(mpg)

c + geom_area(stat = "bin")
x, y, alpha, color, fill, linetype, size

c + geom_density(kernel = "gaussian"**)**
x, y, alpha, color, fill, group, linetype, size, weight

c + geom_dotplot()
x, y, alpha, color, fill

c + geom_freqpoly() x, y, alpha, color, group, linetype, size

c + geom_histogram(binwidth = 5**)** x, y, alpha, color, fill, linetype, size, weight

TWO VARIABLES

continuous bivariate distribution

h <- ggplot(diamonds, aes(carat, price))

h + geom_bin2d(binwidth = c(0.25, 500)**)**
x, y, alpha, color, fill, linetype, size, weight

h + geom_density2d()
x, y, alpha, colour, group, linetype, size

h + geom_hex()
x, y, alpha, colour, fill, size

THREE VARIABLES

seals$z <- with(seals, sqrt(delta_long^2 + delta_lat^2))l <- ggplot(seals, aes(long, lat))

l + geom_contour(aes(z = z))
x, y, z, alpha, colour, group, linetype, size, weight

l + geom_raster(aes(fill = z), hjust=0.5, vjust=0.5, interpolate=FALSE)
x, y, alpha, fill

l + geom_tile(aes(fill = z)), x, y, alpha, color, fill, linetype, size, width

谢谢聆听
Thank you

教师个人联系方式

艾新波

手机：13641159546

QQ： 23127789

微信：13641159546

E-mail: 13641159546@126.com

axb@bupt.edu.cn

地址：北京邮电大学科研楼917室

课程网址：https://github.com/byaxb/RDataAnalytics