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

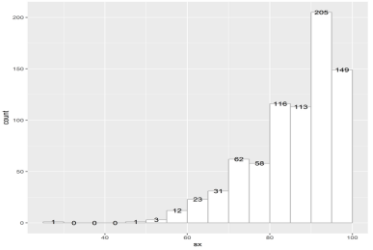
- R语言数据分析
  - 上部：论道 
    - 第1章 气象万千、数以等观
    - 第2章 所谓学习、归类而已
    - 第3章 格言联璧话学习
    - 第4章 源于数学、归于工程
  - 中部：执具 
    - 第5章 工欲善其事必先利其器
    - 第6章 基础编程
    - 第7章 数据对象
  -
    - 第8章 人人都爱tidyverse
    - 第9章 最美不过数据框
  - 下部 博术 
    - 第10章 观数以形
    - 第11章 相随相伴、谓之关联
    - 第12章 既是世间法、自当有分别
    - 第13章 方以类聚、物以群分
 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

一维数据空间形态

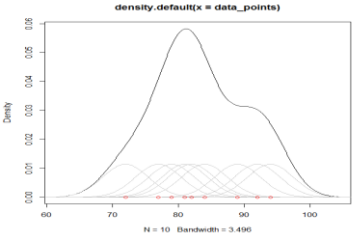
```
The decimal point is at the |

89 | 0
90 | 0
91 | 00
92 | 0000
93 | 0000
94 | 00
95 | 000000
96 | 000000000000000
97 | 0000000000
98 | 000
99 | 0
```

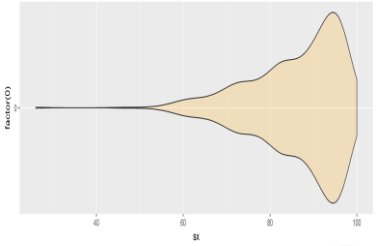
茎叶图



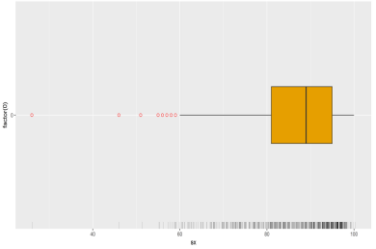
直方图



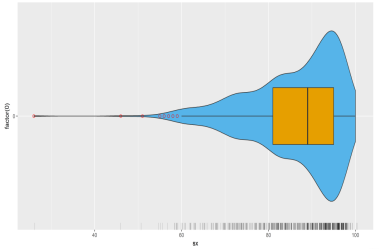
概率密度图



小提琴图



箱线图

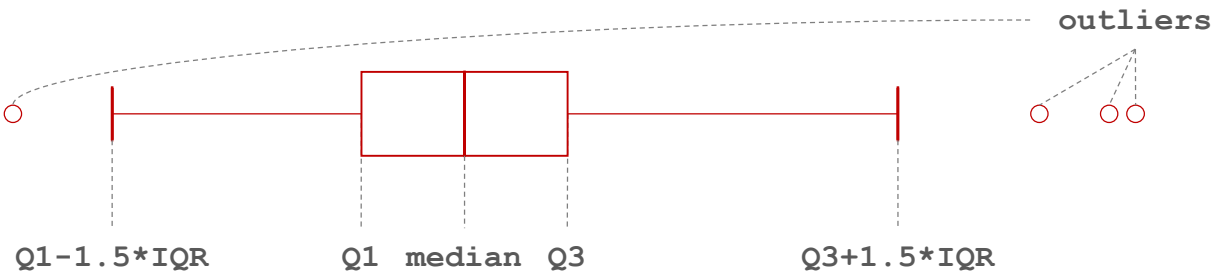


复合图形

箱线图

箱线图通过分位数来刻画数据的分布

如：集中趋势、分散程度、分布形状、异常数据等



Median：有一半的数数小于这个数

Q1:有25%的数小于这个数

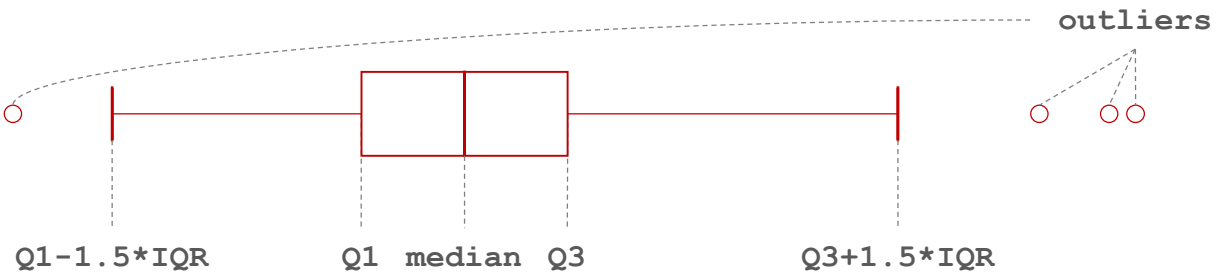
Q3：有75%的数小于这个数

IQR：Inter-Quantile Range，Q3-Q1

箱线图

箱线图通过分位数来刻画数据的分布

如：集中趋势、分散程度、分布形状、异常数据等



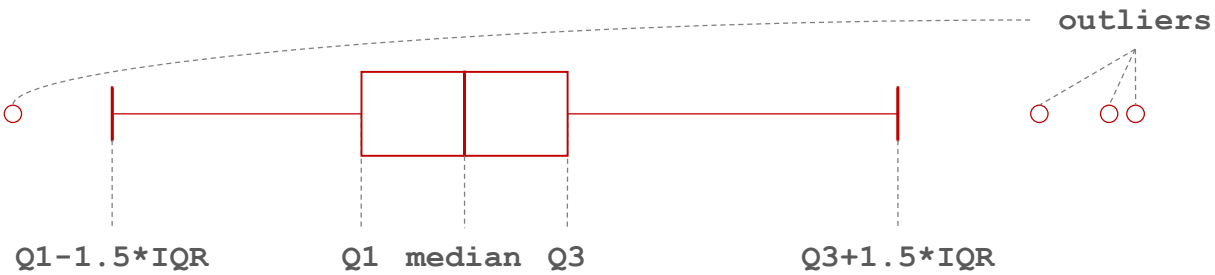
上边界upper whiskers： min{max(x), Q3+1.5×IQR}

下边界lower whiskers： max{min(x), Q1−1.5×IQR}

箱线图

箱线图通过分位数来刻画数据的分布

如：集中趋势、分散程度、分布形状、异常数据等



集中趋势：中位数，大致的平均水平

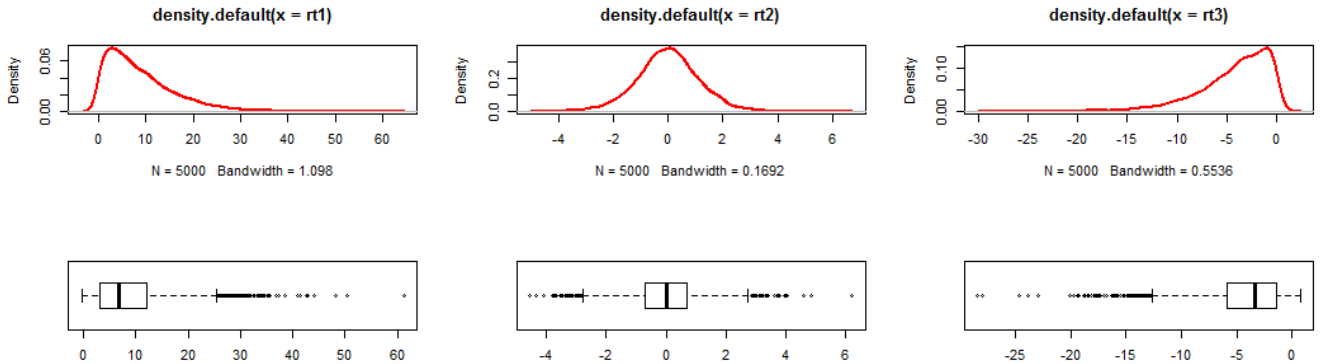
分散程度：IQR，箱子的长度

异常数据：超过上边界或下边界的数据

箱线图

箱线图通过分位数来刻画数据的分布

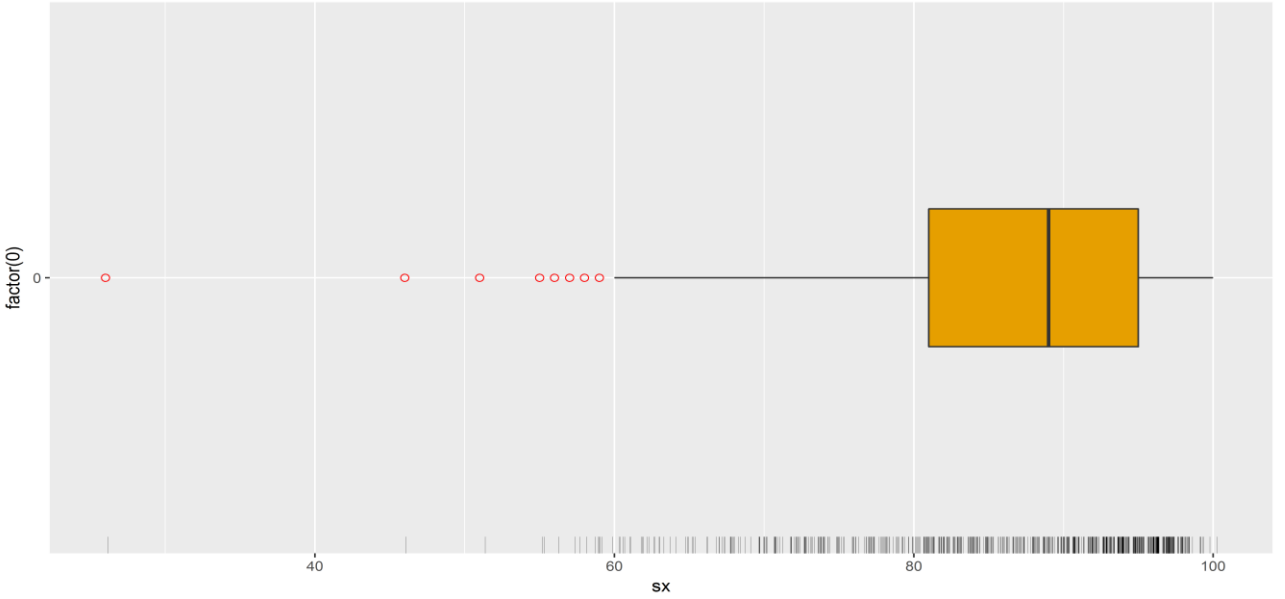
如：集中趋势、分散程度、分布形状、异常数据等



箱线图的R语言实现

```
cjb %>%
  ggplot(aes(x = factor(0), y = sx)) +
  geom_boxplot(width = 0.25,
               fill = "#E69F00",
               outlier.colour = "red",
               outlier.shape = 1,
               outlier.size = 2)+
  geom_rug(position = "jitter",
           size = 0.1,
           sides = "b") +
  coord_flip()
```

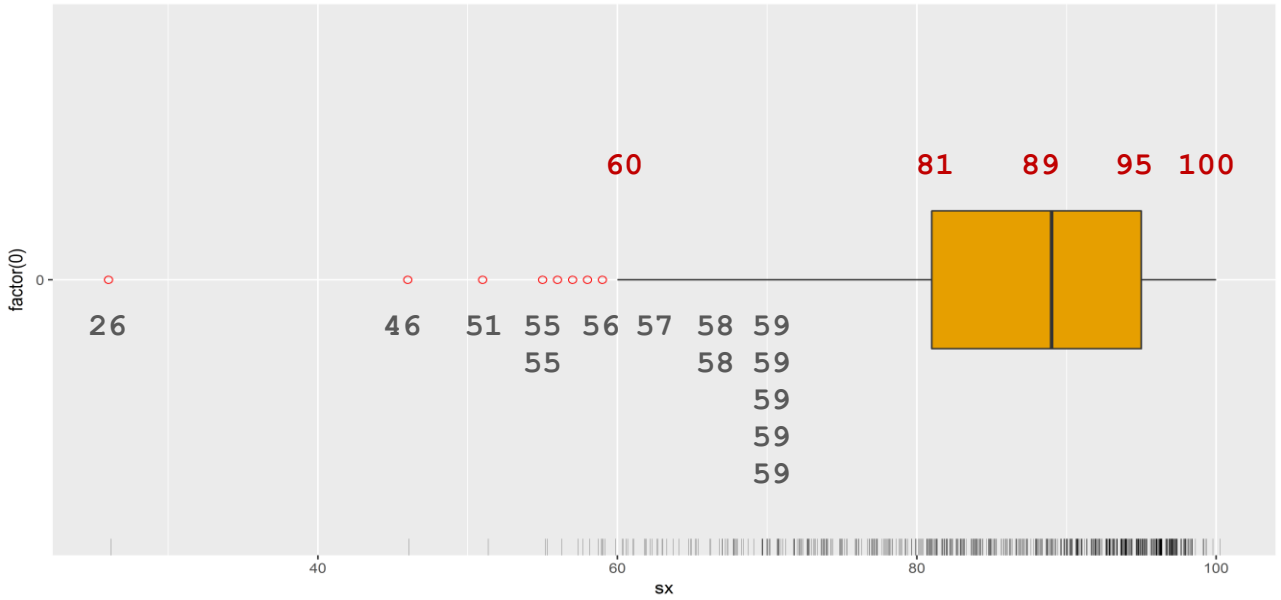
箱线图的R语言实现



箱线图的R语言实现

```
#箱线图一些具体指标
boxplot.stats(cjb$sx)
#> $`stats`
#> [1]  60  81  89  95 100
#> $n
#> [1] 774
#> $conf
#> [1] 88.20491 89.79509
#> $out
#> [1] 55 59 57 59 58 51 56 55 59 26 58 46 59 59
```

箱线图的R语言实现

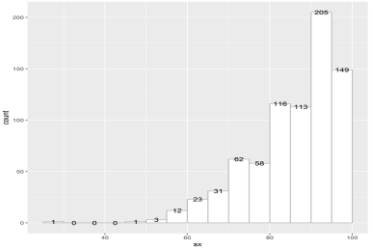

60
81
89
95
100
factor(0)
0
26
46
51
55
55
56
57
58
58
59
59
59
59
59
40
60
80
100
sx

一维数据空间形态

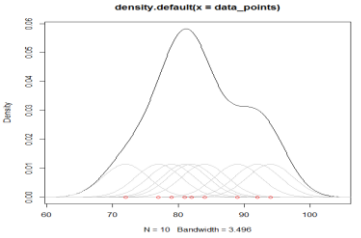
```
The decimal point is at the |

89 | 0
90 | 0
91 | 00
92 | 0000
93 | 0000
94 | 00
95 | 000000
96 | 0000000000000000
97 | 0000000000
98 | 000
99 | 0
```

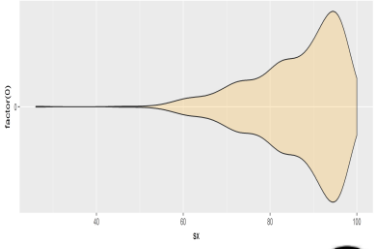
茎叶图



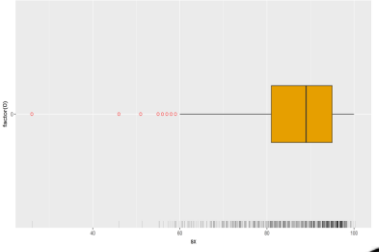
直方图



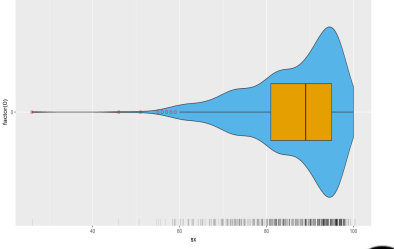
概率密度图



小提琴图



箱线图

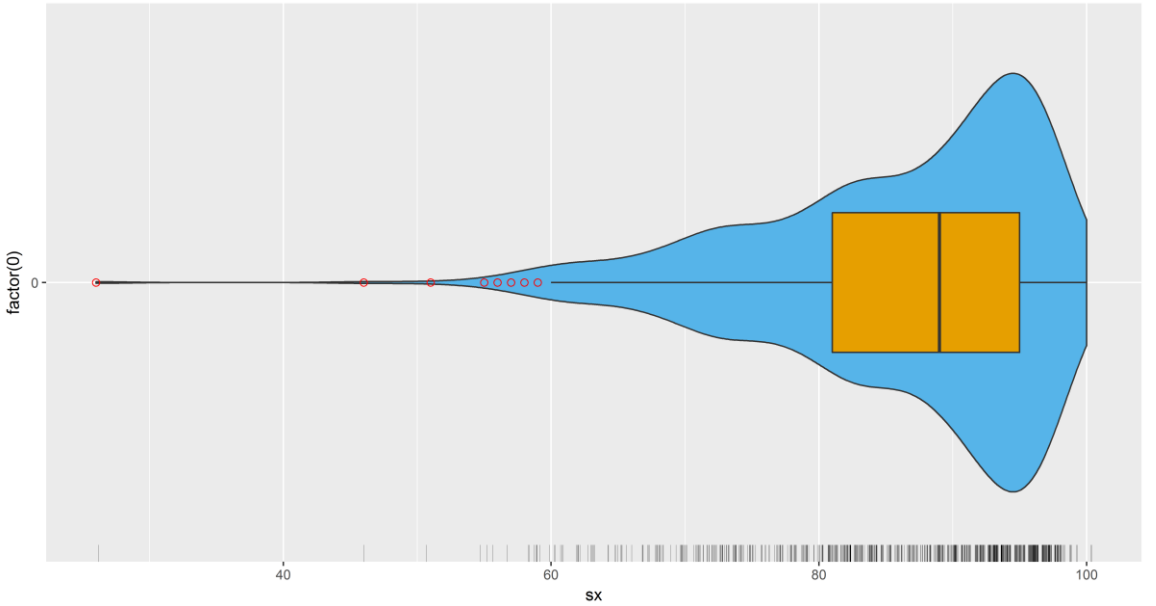


复合图形

箱线图+小提琴图

```
cjb %>%
  ggplot(aes(x = factor(0), y = sx)) +
  geom_violin(fill = "#56B4E9", width = 0.75) +
  geom_boxplot(width = 0.25,
               fill = "#E69F00",
               outlier.colour = "red",
               outlier.shape = 1,
               outlier.size = 2)+
  geom_rug(position = "jitter",
           size = 0.1,
           sides = "b") +
  coord_flip()
```

箱线图+小提琴图

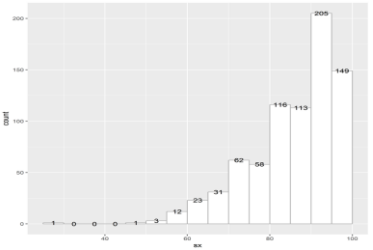


一维数据空间形态

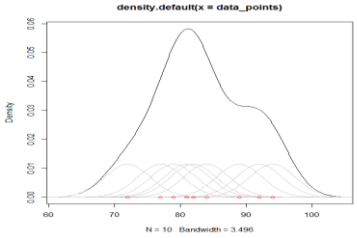
```
The decimal point is at the |

 89 | 0
 90 | 0
 91 | 00
 92 | 0000
 93 | 0000
 94 | 00
 95 | 000000
 96 | 0000000000000000
 97 | 0000000000
 98 | 000
 99 | 0
```

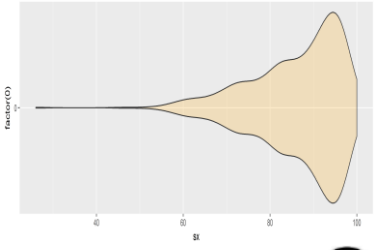
茎叶图



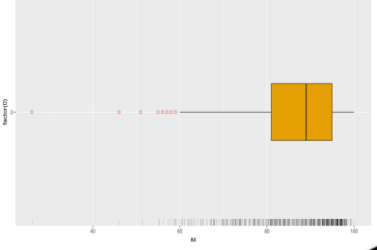
直方图



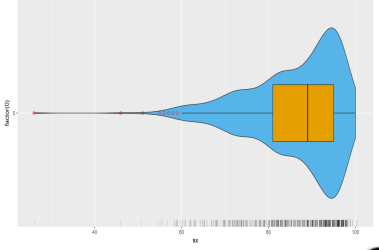
概率密度图



小提琴图



箱线图



复合图形

定量刻画：集中趋势

平均值：

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

平均值包含了每个数据点的信息，但是容易受异常值影响

PS： 平均值不只是反映了数据的整体水平、集中趋势，更是预测和判断时，没有办法的办法

有生于无： 没有模型时，平均值（或众数）是最常见的预测手段

定量刻画：集中趋势

中位数：

将所有数据从小到大进行排列，站在中间的那个数就是中位数

左侧没有那个数比这个数更大，右侧没有哪个数比这个数更小

$$median(x) = \frac{x_{\lceil \frac{n}{2} \rceil} + x_{\lceil \frac{n+1}{2} \rceil}}{2}$$

中位数相对比较稳定，不易受异常值的影响

集中趋势的R语言实现

```
cjb %>%
  group_by(wlfk) %>%#按文理分科分组统计
  summarise(
    count = n(), #各组人数
    sx_median = median(sx),#中位数
    sx_mean = mean(sx))#均值
#> # A tibble: 2 x 4
#>   Wlfk    count  sx_median  sx_mean
#>   <fct>  <int>  <dbl>       <dbl>
#> 1理科     380    93       89.8
#> 2文科     394    84       82.7
```

定量刻画：分散程度

极差range：

$$R = \max(x) - \min(x)$$

四分位距Interquartile range：

$$IQR = Q3 - Q1$$

标准差Standard deviation：

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

分散程度的R语言实现

```
cjb %>%
  group_by(wlfk) %>%#按文理分科分组统计
  summarise(
    sx_max = max(sx), #最大值
    sx_min = min(sx),#最小值
    sx_range = max(sx) - min(sx))#极差
#> # A tibble: 2 x 4
#>   wlfk  sx_max sx_min sx_range
#>   <fct>  <dbl>  <dbl>    <dbl>
#> 1 理科     100     46       54
#> 2 文科     100     26       74
```

分散程度的R语言实现

```
cjb %>%
  group_by(wlfk) %>%#按文理分科分组统计
  summarise(
    sx_Q3 = quantile(sx, 3/4),#第三分位数
    sx_Q1 = quantile(sx, 1/4),#第一分位数
    sx_iqr = IQR(sx))#四分位距
#> # A tibble: 2 x 4
#>   wlfk  sx_Q3 sx_Q1 sx_iqr
#>   <fct> <dbl> <dbl>  <dbl>
#> 1 理科     96    86     10
#> 2 文科     92    75     17
```

求各科的分布指标

```
#查看各科情况
round(apply(cjb[, 4:12], 2, function(x) {
  c(mean = mean(x),
    median = median(x),
    range = diff(range(x)),
    IQR = IQR(x))
}))
#>        yw sx wy zz  ls dl wl hx sw
#> mean   87 86 88 92  89 93 81 92 86
#> median 88 89 88 93  90 94 83 94 88
#> range  63 74 69 35 100 30 79 48 45
#> IQR     6 14  8  5  10  6 17 10 12
```

单变量（一维数据空间）的散布形态

可视化类型	数据分布形态	大量数据	保留原始信息	展示分位数	异常情况
茎叶图	☑	☐	☑	☐	☐
直方图	☑	☑	☐	☐	☐
概率密度图	☑	☑	☐	☐	☐
小提琴图	☑	☑	☐	☐	☐
箱线图	☑	☑	☐	☑	☑

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