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

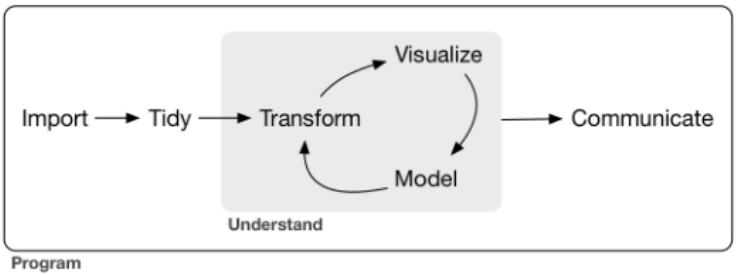
艾新波 / 2018 • 北京

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

一个完整的数据分析过程

CRISP



Hadley: R for Data Science

观数以形：未知其内、先表其外

数据既不是物质，也不是能量

那么

它长什么样呢？

刘徽：析理以辞，解体**用图**

柏拉图：上帝终究要将世界**几何化**

数与形

数形本是相倚依，焉能分作两边飞

数缺形时少直观，形缺数时难入微

数形结合百般好，隔裂分家万事非

几何代数统一体，永远联系莫分离

该诗来自网络

与《谈谈与蜂房结构有关的数学问题（华罗庚）》中的表述略有不同

机器学习的核心：

关系结构

机器所能学到的主要是

变量之间的**关系**和数据空间的**结构**

认识数据的方法

认识数据的核心依然在于：

关系结构

主要是数据空间

所呈现的几何**形态**及并用少量数据予以**量化**

数据空间的结构和形态

空间

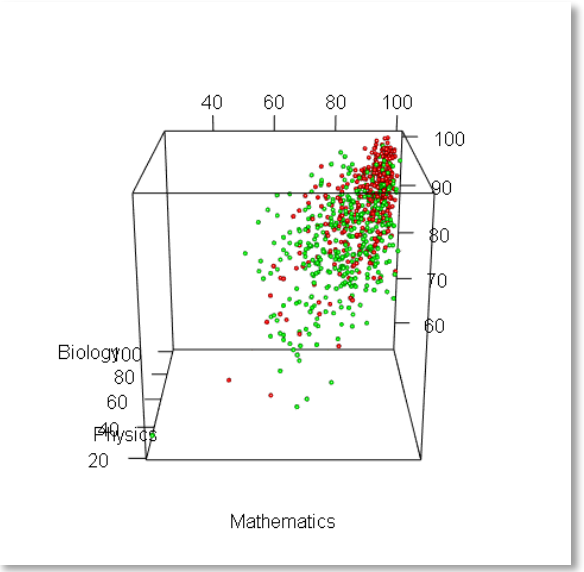
本质就是集合

维度

列/字段/变量/属性/特征

数据点

行/特征向量/有序数对

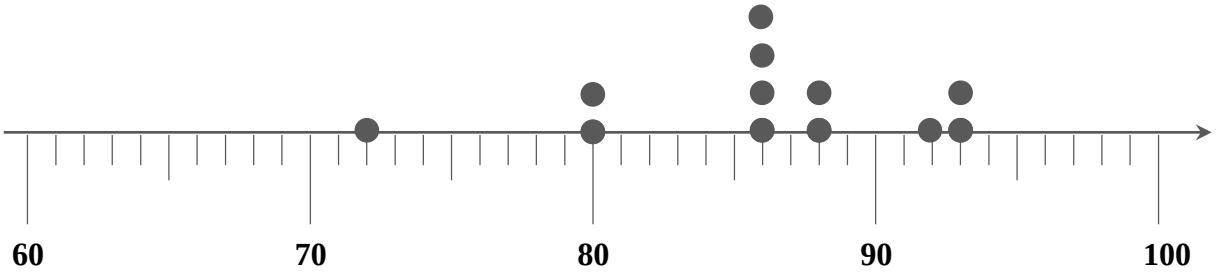


认识数据 ≈

通过几何的方式看数据

观数以形，辨形以识数

一维数据空间形态

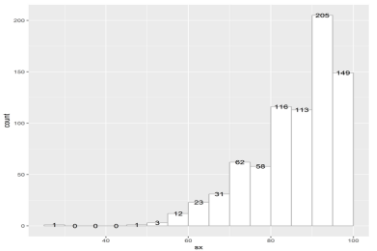


一维数据空间形态

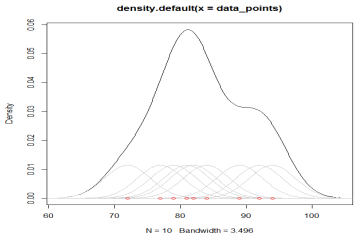
```
The decimal point is at the |

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

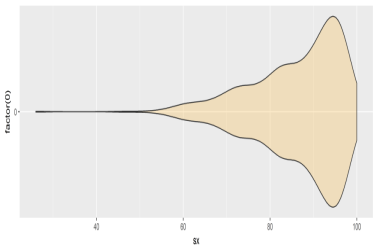
茎叶图



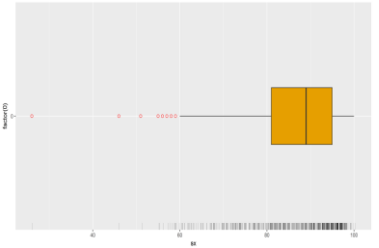
直方图



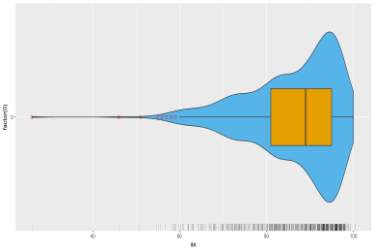
概率密度图



小提琴图



箱线图



复合图形

茎叶图

1101班所有同学的数学成绩（52名同学）

82 94 79 84 92 82 72 89 77 81 83 82 84 81 85 84 71 84
78 79 61 74 88 70 55 92 78 83 78 74 70 70 73 65 74 72
64 71 81 74 70 69 73 60 65 68 71 60 59 57 59 67

stem	leaf
5	5799
6	001455789
7	000011122334444788899
8	111222334444589
9	224

$$value \approx stem \times 10^{n} + leaf \times 10^{n-1}$$
$$n = 1$$

茎叶图

1101班所有同学的数学成绩（52名同学）

55 57 59 59 60 60 61 64 65 65 67 68 69 70 70 70 70 71
71 71 72 72 73 73 74 74 74 74 77 78 78 78 79 79 81 81
81 82 82 82 83 83 84 84 84 84 85 88 89 92 92 94

stem	leaf
5	5799
6	001455789
7	000011122334444788899
8	111222334444589
9	224

$$value \approx stem \times 10^{n} + leaf \times 10^{n-1}$$
$$n = 1$$

茎叶图

542.2 564.2 590.9 593.7 596.6 606.0 612.6 637.7 643.7 654.1
672.8 673.1 681.6 688.2 691.8 692.5 703.7 718.9 719.3 724.3
724.4 729.8 731.4 731.5 736.3 742.4 749.4 760.3 767.9 778.6
779.8 789.4 789.9 797.7 801.4 801.8 812.5 819.5 822.6 825.4
825.6 828.6 831.8 840.0 840.4 845.5 853.5 865.7 891.0 912.6
915.9 950.0

stem	leaf
5	4699
6	011445778999
7	02222333445678899
8	00012233334455 79
9	125

$$value \approx stem \times 10^{n} + leaf \times 10^{n-1}$$
$$n = 2$$

茎叶图

```r
#1101班数学成绩茎叶图
cjb %>%
  filter(bj == "1101") %>%
  select(sx) %>%
  as_vector() %>%
  stem(scale = 0.5)
#1110班数学成绩茎叶图
cjb%>%
  filter(bj == "1110") %>%
  select(sx) %>%
  as_vector() %>%
  stem(scale = 2)
```

```r
#1101班数学成绩茎叶图
cjb %>%
  filter(bj == "1101") %>%
  pull(sx) %>%
  stem(scale = 0.5)

#1110班数学成绩茎叶图
cjb%>%
  filter(bj == "1110") %>%
  pull(sx) %>%
  stem(scale = 2)
```

茎叶图的R语言实现

The decimal point is 1 digit(s) to the right of the |

5 | 5799
6 | 001455789
7 | 000011122334444788899
8 | 111222334444589
9 | 224

1101班（文科）

The decimal point is at the |

89 | 0
90 | 0
91 | 00
92 | 0000
93 | 0000
94 | 00
95 | 000000
96 | 000000000000000
97 | 000000000
98 | 000
99 | 0

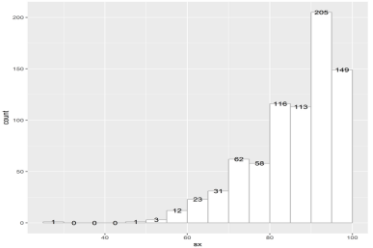
1110班（理科）

一维数据空间形态

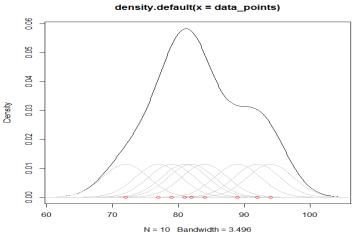
```
The decimal point is at the |

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

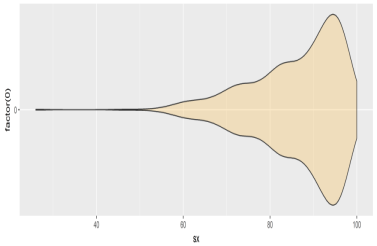
茎叶图



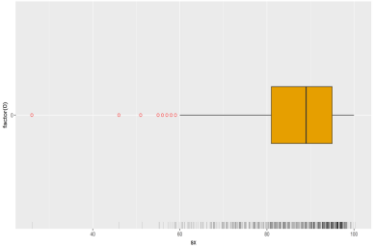
直方图



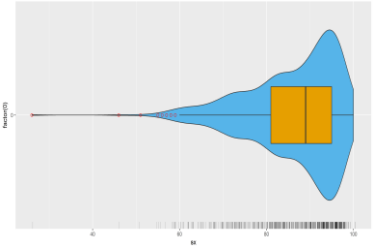
概率密度图



小提琴图



箱线图

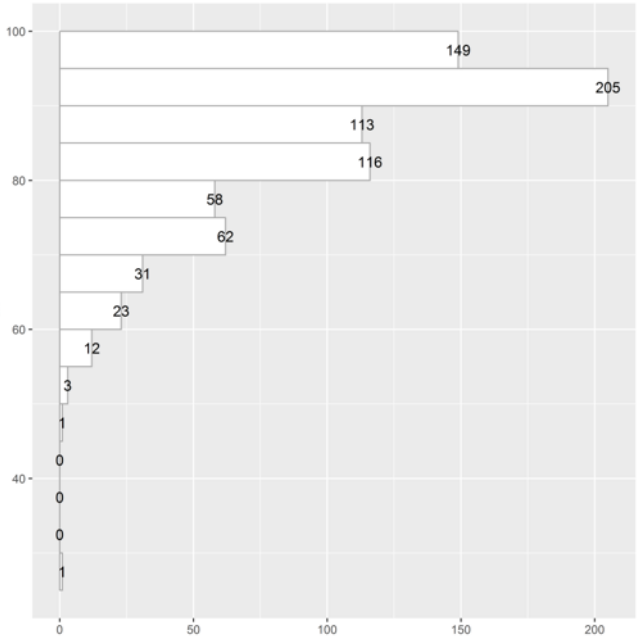


复合图形

直方图

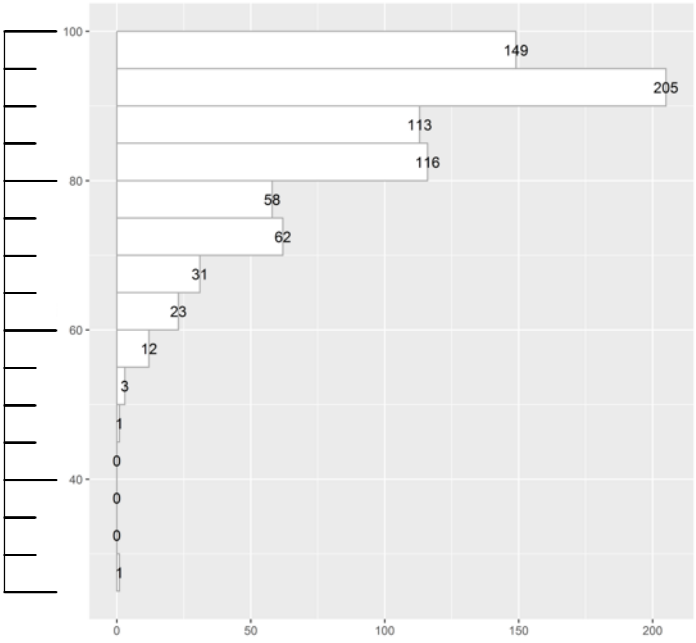
The decimal point is at the |

89 | 0
90 | 0
91 | 00
92 | 0000
93 | 0000
94 | 00
95 | 000000
96 | 000000000000000
97 | 000000000
98 | 000
99 | 0



直方图

i	bin	$count_i$
15	(95,100]	149
14	(90,95]	205
13	(85,90]	113
12	(80,85]	116
11	(75,80]	58
10	(70,75]	62
9	(65,70]	31
8	(60,65]	23
7	(55,60]	12
6	(50,55]	3
5	(45,50]	1
4	(40,45]	0
3	(35,40]	0
2	(30,35]	0
1	(25,30]	1



直方图

i	bin	$count_i$
15	(95,100]	149
14	(90,95]	205
13	(85,90]	113
12	(80,85]	116
11	(75,80]	58
10	(70,75]	62
9	(65,70]	31
8	(60,65]	23
7	(55,60]	12
6	(50,55]	3
5	(45,50]	1
4	(40,45]	0
3	(35,40]	0
2	(30,35]	0
1	(25,30]	1

$$n = \sum_{i=1}^{k} count_i$$

k可通过以下方法确定：

Sturges: $k = \lceil log_2 n \rceil + 1$

$$k = \left\lceil \frac{\max(x) - \min(x)}{h} \right\rceil$$

Scott: $h = \frac{3.5\hat{\sigma}}{\frac{1}{n^3}}$

Freedman–Diaconis: $h = 2\frac{IQR(x)}{\frac{1}{n^3}}$

直方图的R语言实现

```
#看一看数据分布的形状
sx_hist_results <- hist(cjb$sx, plot = FALSE)
#查看sx_hist_results的类型
typeof(sx_hist_results)
#> [1] "list"
#查看列表的组成
names(sx_hist_results)
#> [1] "breaks"   "counts"   "density"  "mids"
"xname"    "equidist"
```

直方图的R语言实现

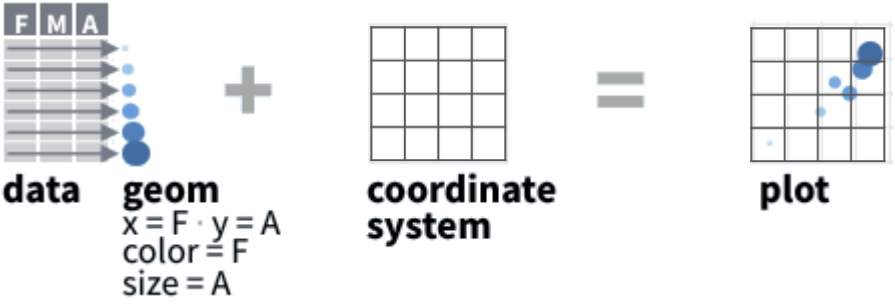
#绘制直方图

```
ggplot(data = cjb, mapping = aes(sx)) +
  geom_histogram(
    breaks = sx_hist_results$breaks,
    color = "darkgray",
    fill = "white") +
  stat_bin(breaks = sx_hist_results$breaks,
           geom = "text",
           aes(label = ..count..)) +
  coord_flip()
```

初探ggplot2：cheatsheet

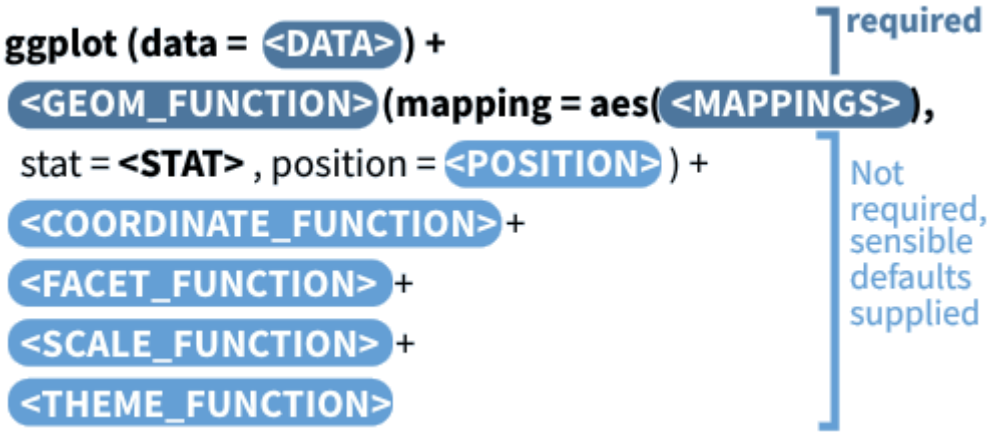
ggplot2 is based on the grammar of graphics, the idea that you can build every graph from the same components: a data set, a coordinate system, and geoms -visual marks that represent data points.

To display values, map variables in the data to visual properties of the geom (aesthetics) like size, color, and x and y locations.



初探ggplot2：cheatsheet

Complete the template below to build a graph:



更多内容请参阅：Data Visualization with ggplot2 : : CHEAT SHEET@RStudio 以及R for Data Science: http://r4ds.had.co.nz/ （☆强烈推荐☆）

直方图的R语言实现

#绘制直方图

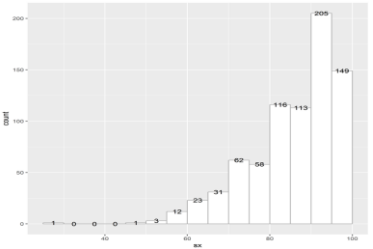
```
ggplot(data = cjb, mapping = aes(sx)) +
  geom_histogram(
    breaks = sx_hist_results$breaks,
    color = "darkgray",
    fill = "white") +
  stat_bin(breaks = sx_hist_results$breaks,
           geom = "text",
           aes(label = ..count..)) +
  coord_flip()
```

一维数据空间形态

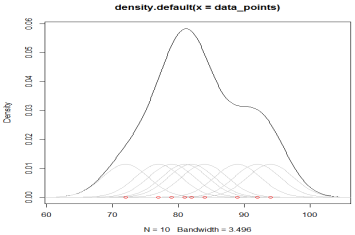
```
The decimal point is at the |

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

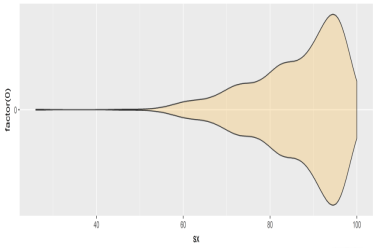
茎叶图



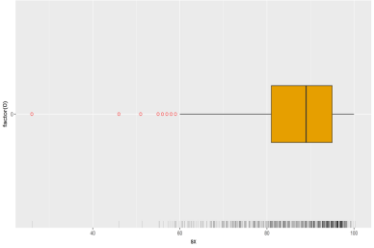
直方图



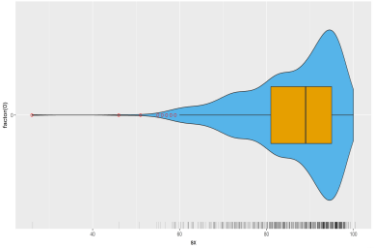
概率密度图



小提琴图

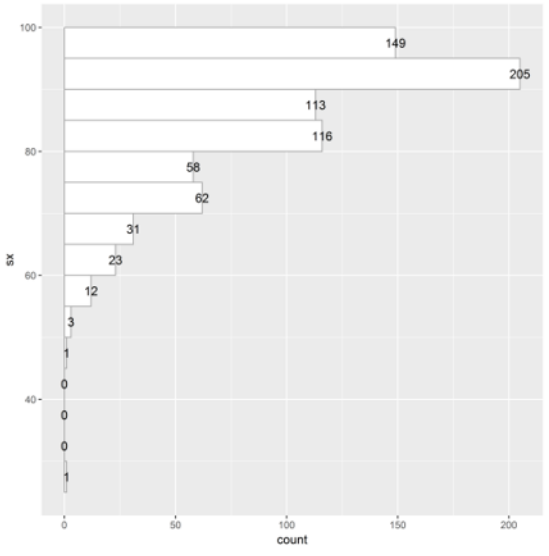


箱线图

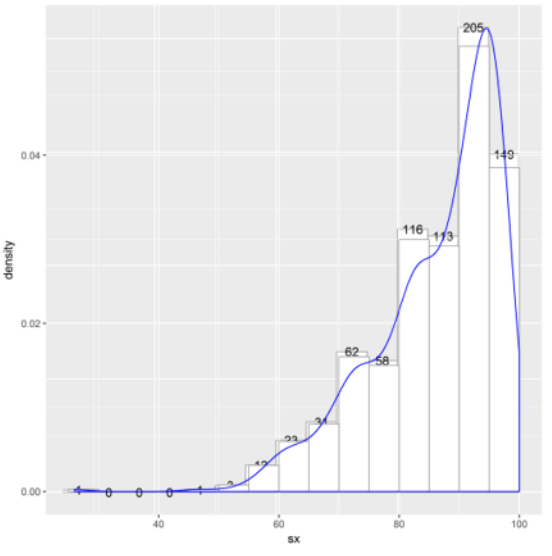


复合图形

直方图的R语言实现



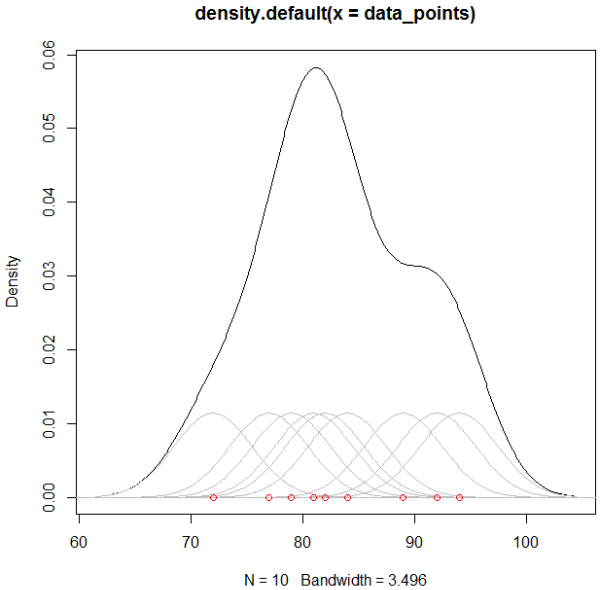
有+coord_flip()



无+coord_flip()

概率密度图

- 概率密度取值大的地方得到样本的可能性更大
- 反之样本分布越密集地方密度函数取值也越大
- 每个样本都对总体概率密度有一定贡献
- 密度函数由总体贡献之和所确定



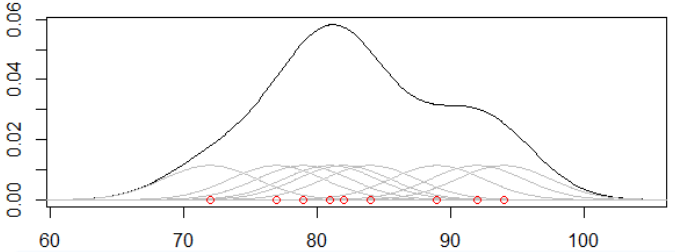
概率密度图

把前述直观判断变为数学语言：

设$x_1, x_2, \cdots, x_n$为变量x的独立同分布的一个样本，则x所服从分布的密度函数的核密度估计为：

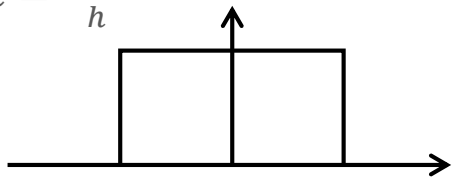
$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K_h\left(\frac{x - x_i}{h}\right)$$

其中$K_h(\cdot)$为核函数，h为窗口宽度

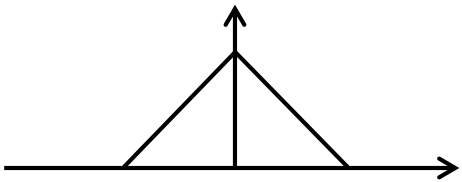


概率密度图

令：$u = \frac{x - x_i}{h}$

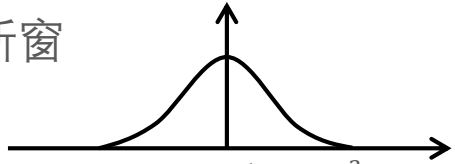


矩形窗 $K(u) = \begin{cases} 1, |u| \le \frac{1}{2} \\ 0, else \end{cases}$



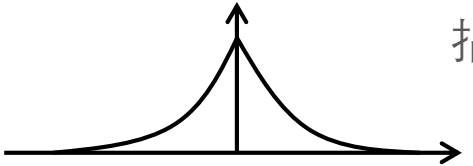
$K(u) = \begin{cases} 1 - |u|, |u| \le 1 \\ 0, else \end{cases}$ 三角窗

高斯窗



$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$

指数窗

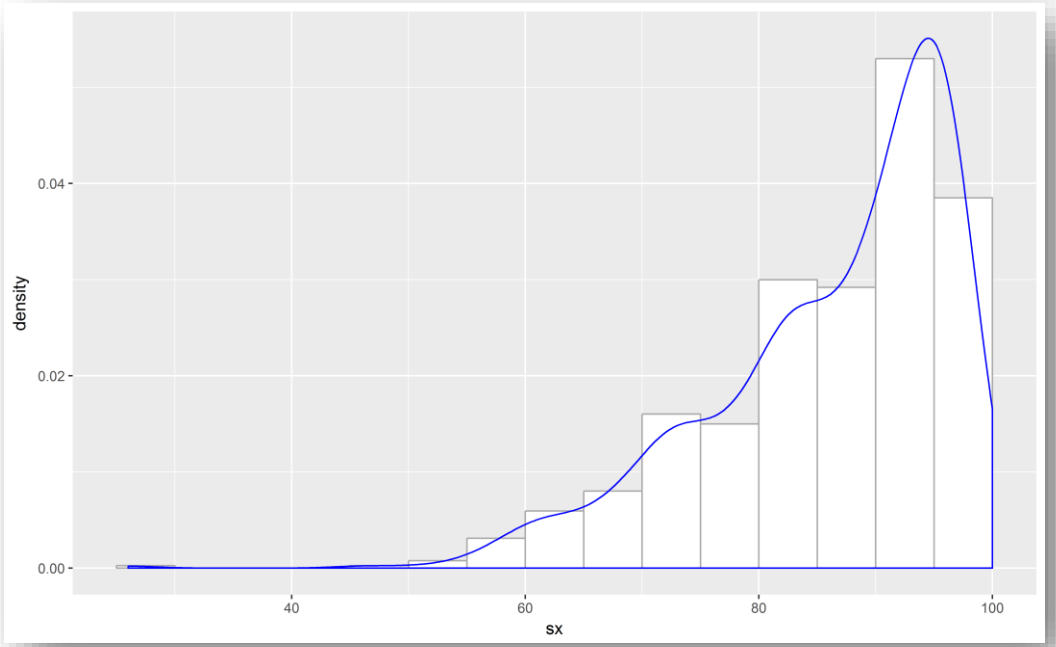


$K(u) = \frac{1}{2} e^{-|u|}$

概率密度图的R语言实现

```
#获取直方图相关参数
sx_hist_results <- hist(cjb$sx,
                        plot = FALSE)
#绘制直方图
ggplot(data = cjb, mapping = aes(sx)) +
  geom_histogram(
    aes(y = ..density..),
    breaks = sx_hist_results$breaks,
    color = "darkgray",
    fill = "white") +
  #绘制概率密度曲线
  geom_density(colour = "blue")
```

概率密度图的R语言实现

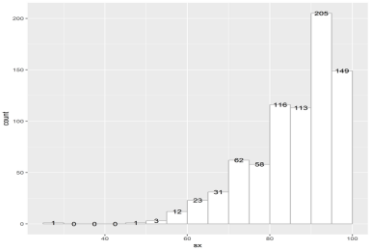


一维数据空间形态

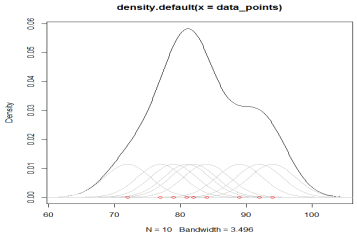
```
The decimal point is at the |

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

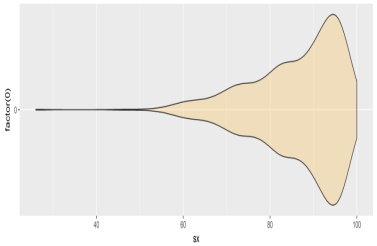
茎叶图



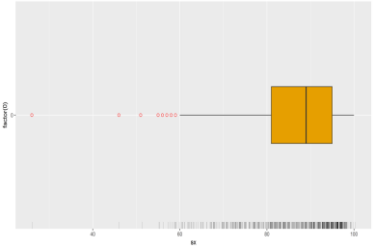
直方图



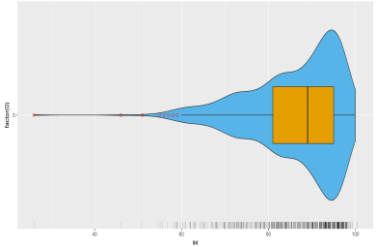
概率密度图



小提琴图



箱线图

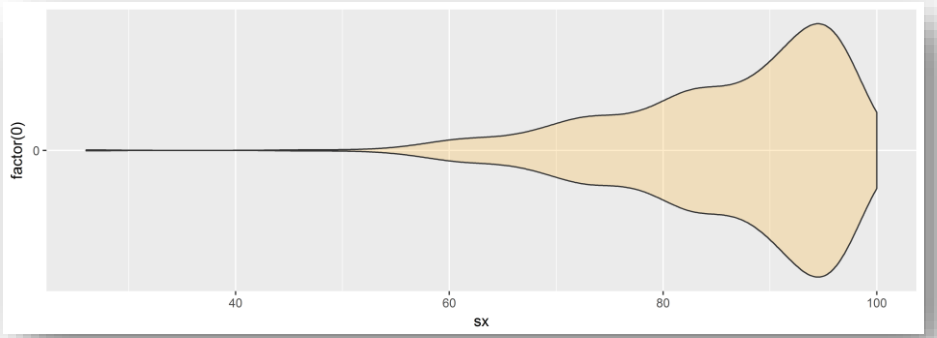


复合图形

小提琴图的R语言实现

#绘制小提琴图

```
ggplot(cjb, aes(x = factor(0), y = sx)) +
  geom_violin(fill = "orange", alpha = 0.2)+
  coord_flip()
```

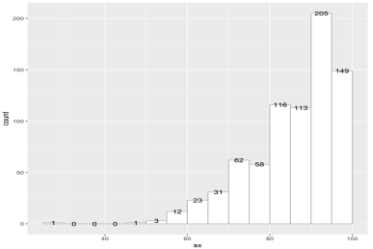


一维数据空间形态

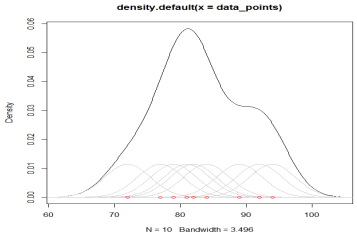
```
The decimal point is at the |

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

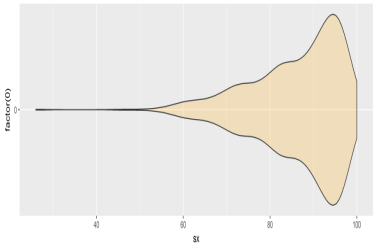
茎叶图



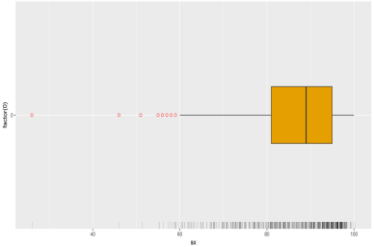
直方图



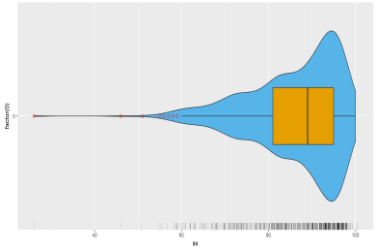
概率密度图



小提琴图



箱线图



复合图形

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