数据科学基础

Foundations of Data Science

5.4 数据可视化

陈振宇
南京大学智能软件工程实验室
www.iselab.cn

一维数据可视化

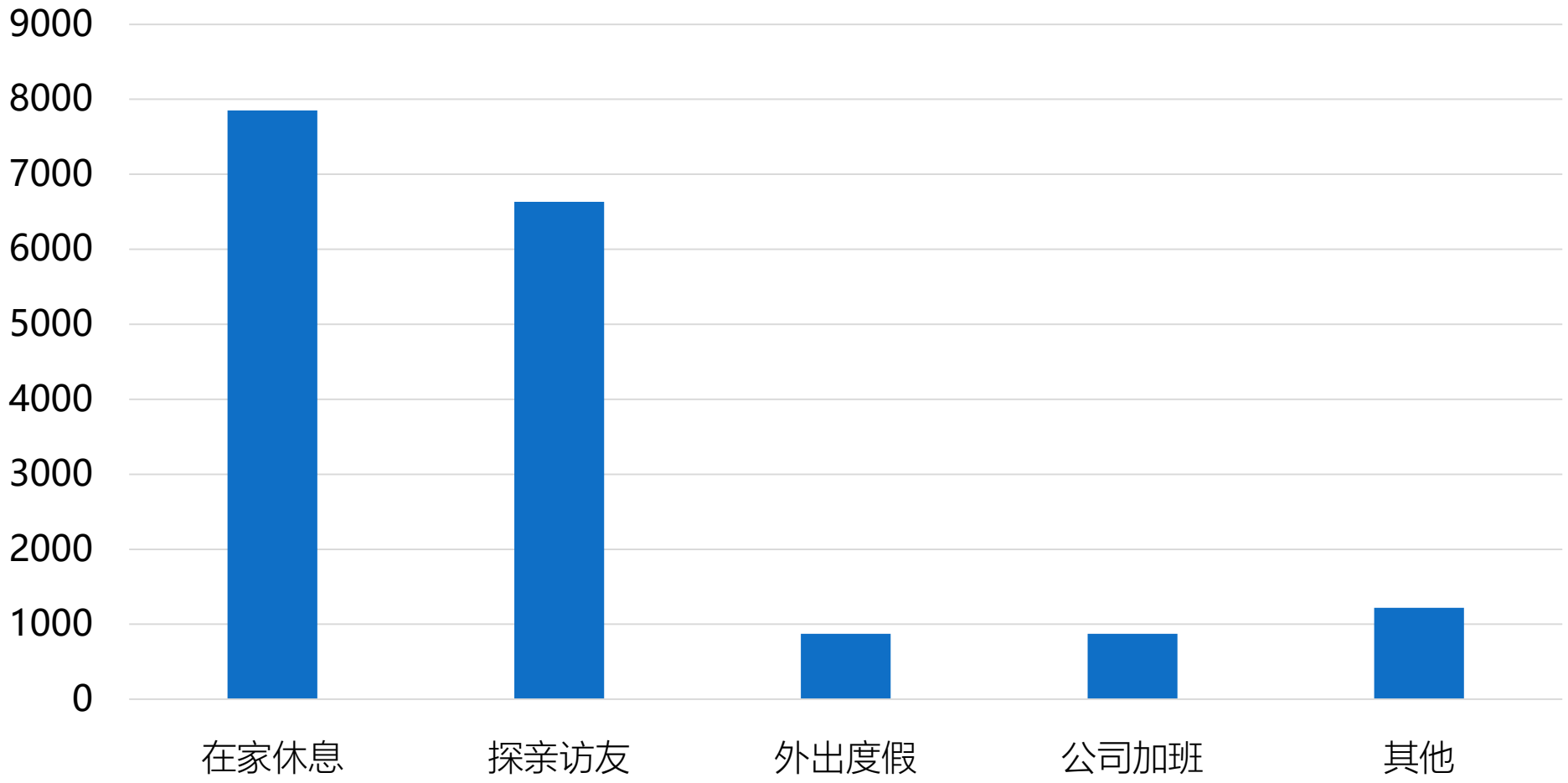
初始数据汇总

- 频数是各个数据被观测到的次数。
- 频率是频数除以总次数。

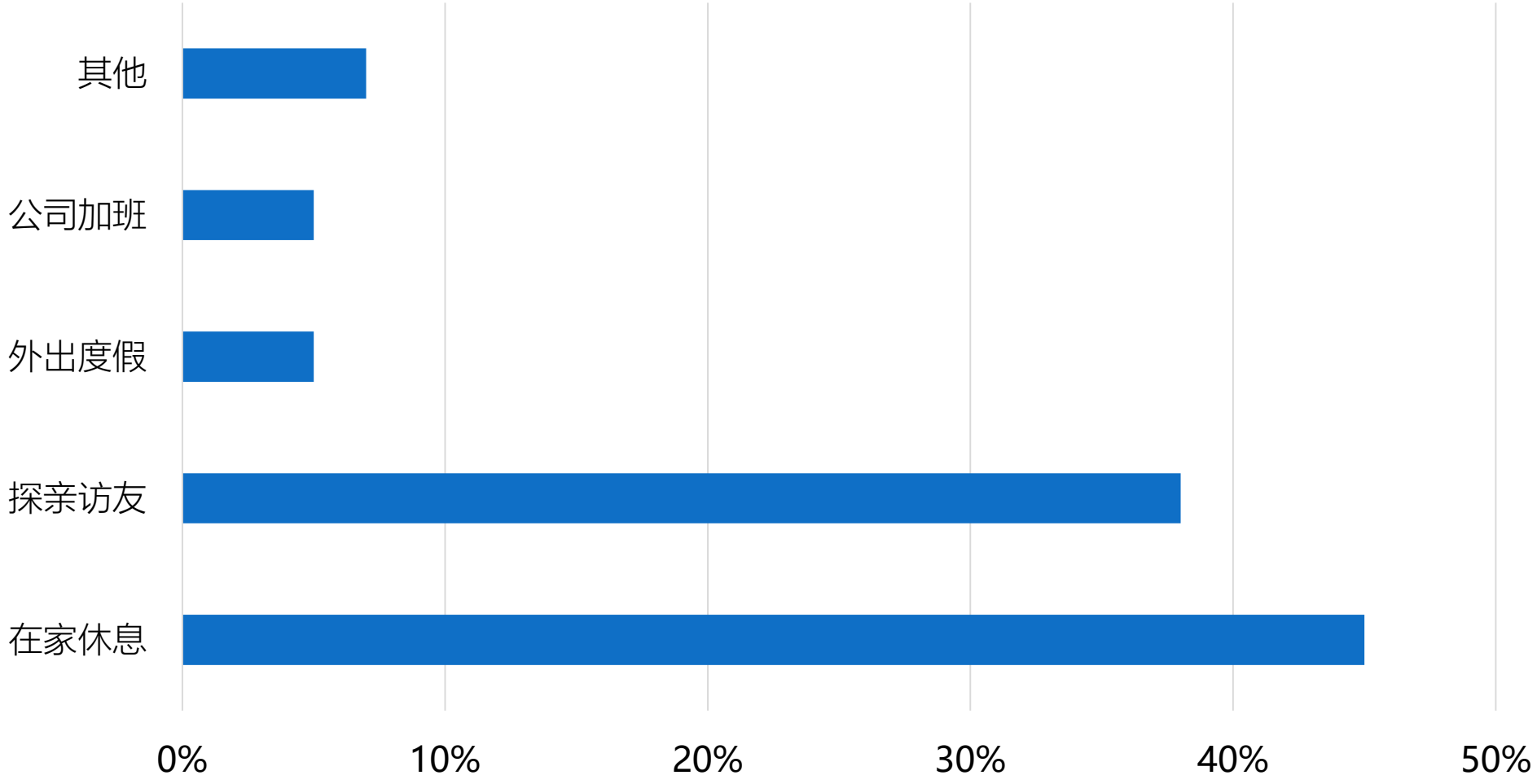
例如，为了进一步改善节假日安排，相关部门进行了一次关于黄金周过节方式的网络调查，收到17452有效调查票。详细调查结果如右边表格。

过节方式	频数	频率
在家休息	7853	0.45
探亲访友	6632	0.38
外出度假	873	0.05
公司加班	873	0.05
其他	1221	0.07
合计	17452	1

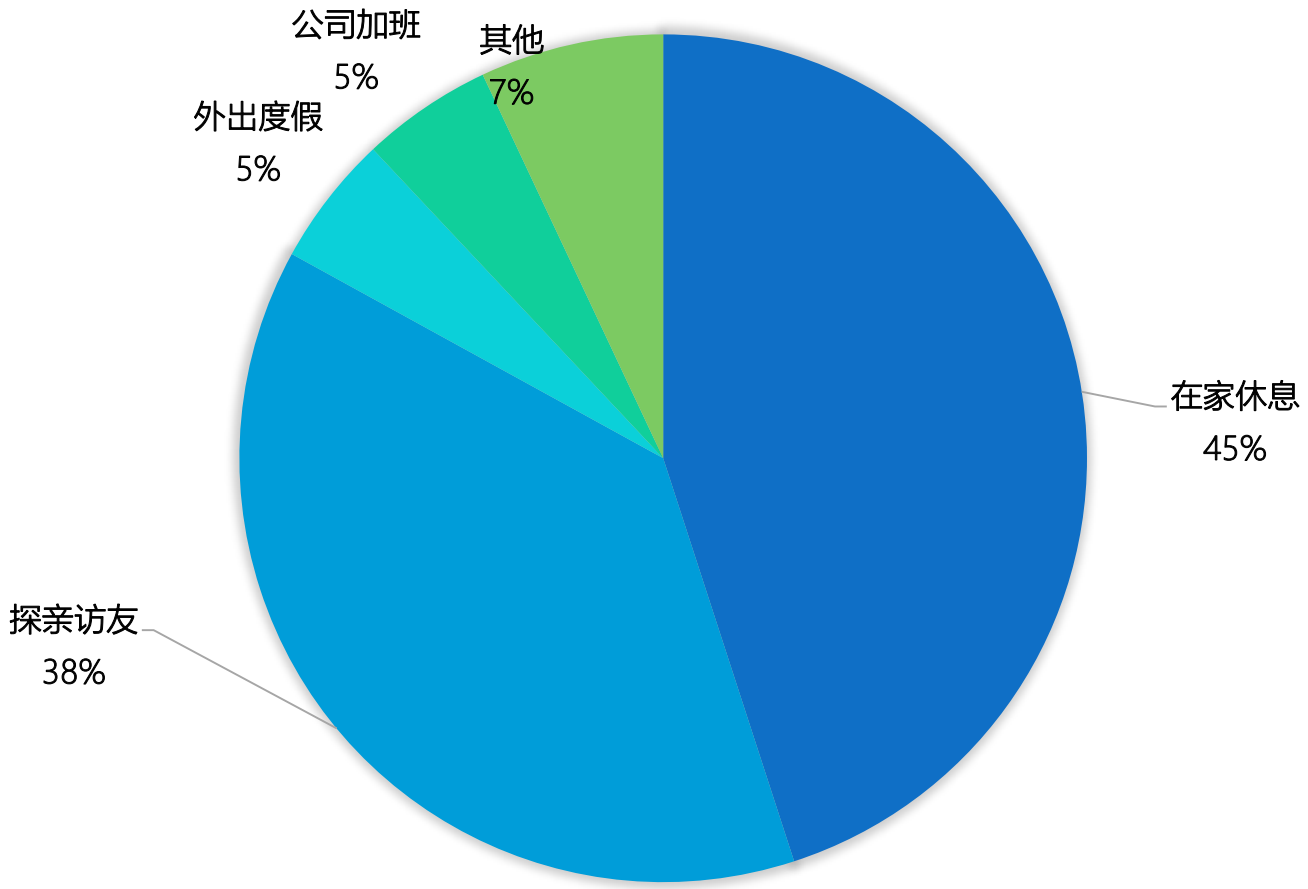
柱状图


9000
8000
7000
6000
5000
4000
3000
2000
1000
0
在家休息
探亲访友
外出度假
公司加班
其他

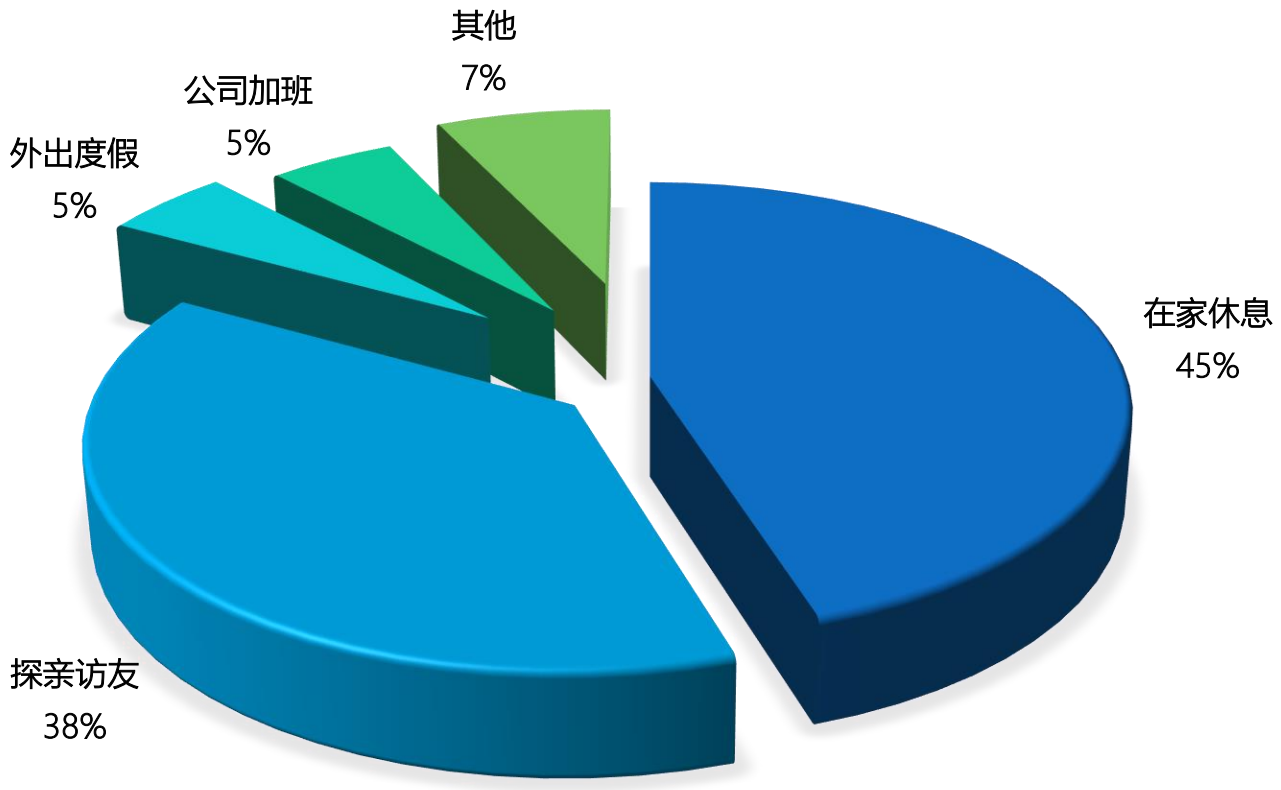
条状图



饼图

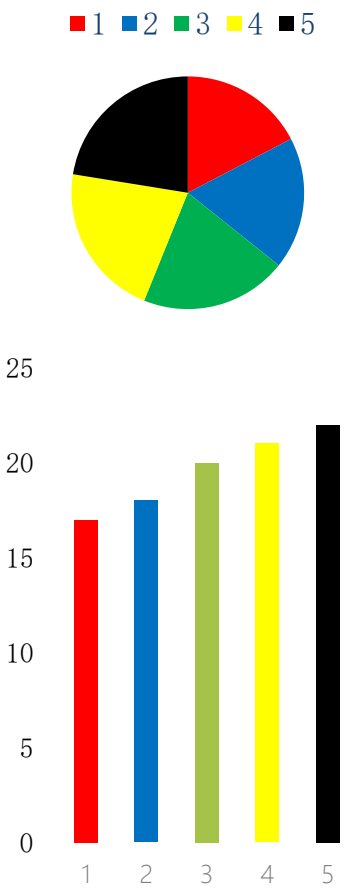


饼图

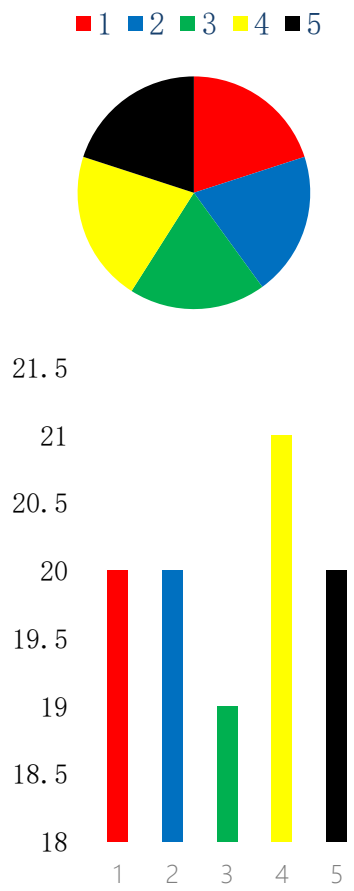


饼图与柱状图对比

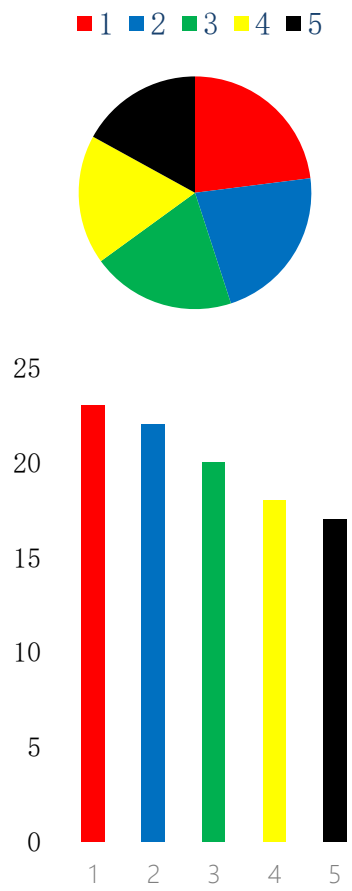
A



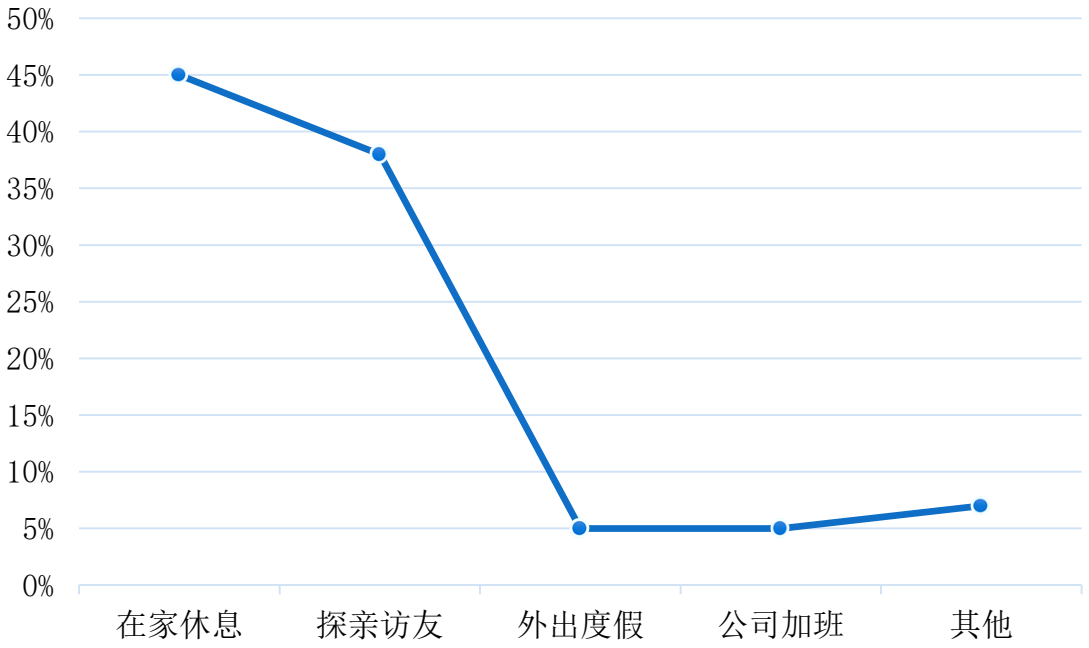
B

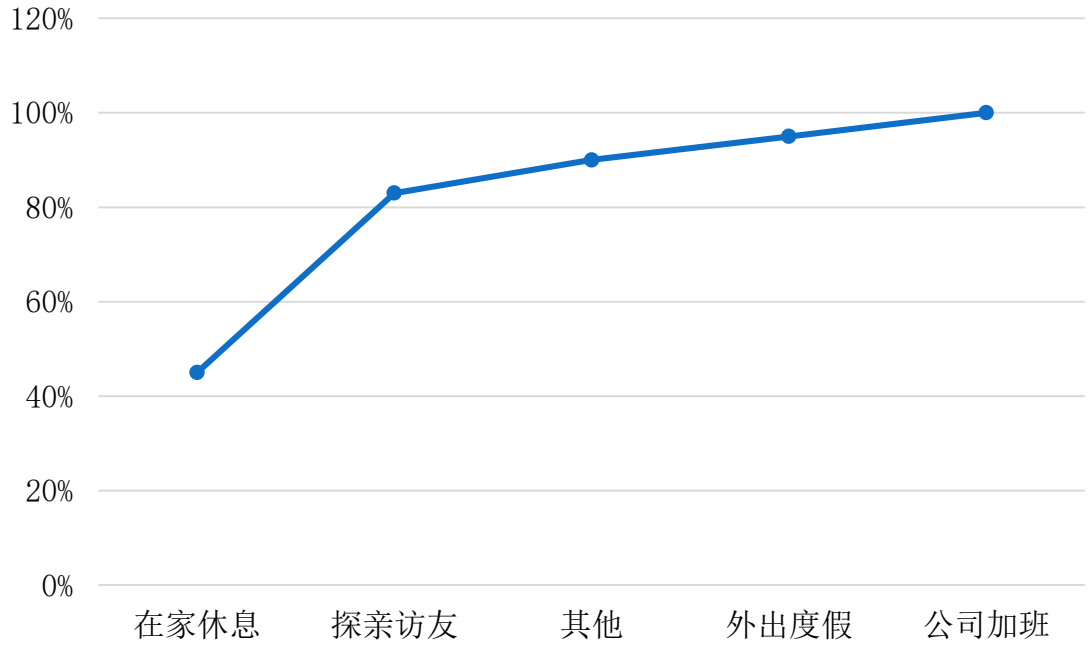


C

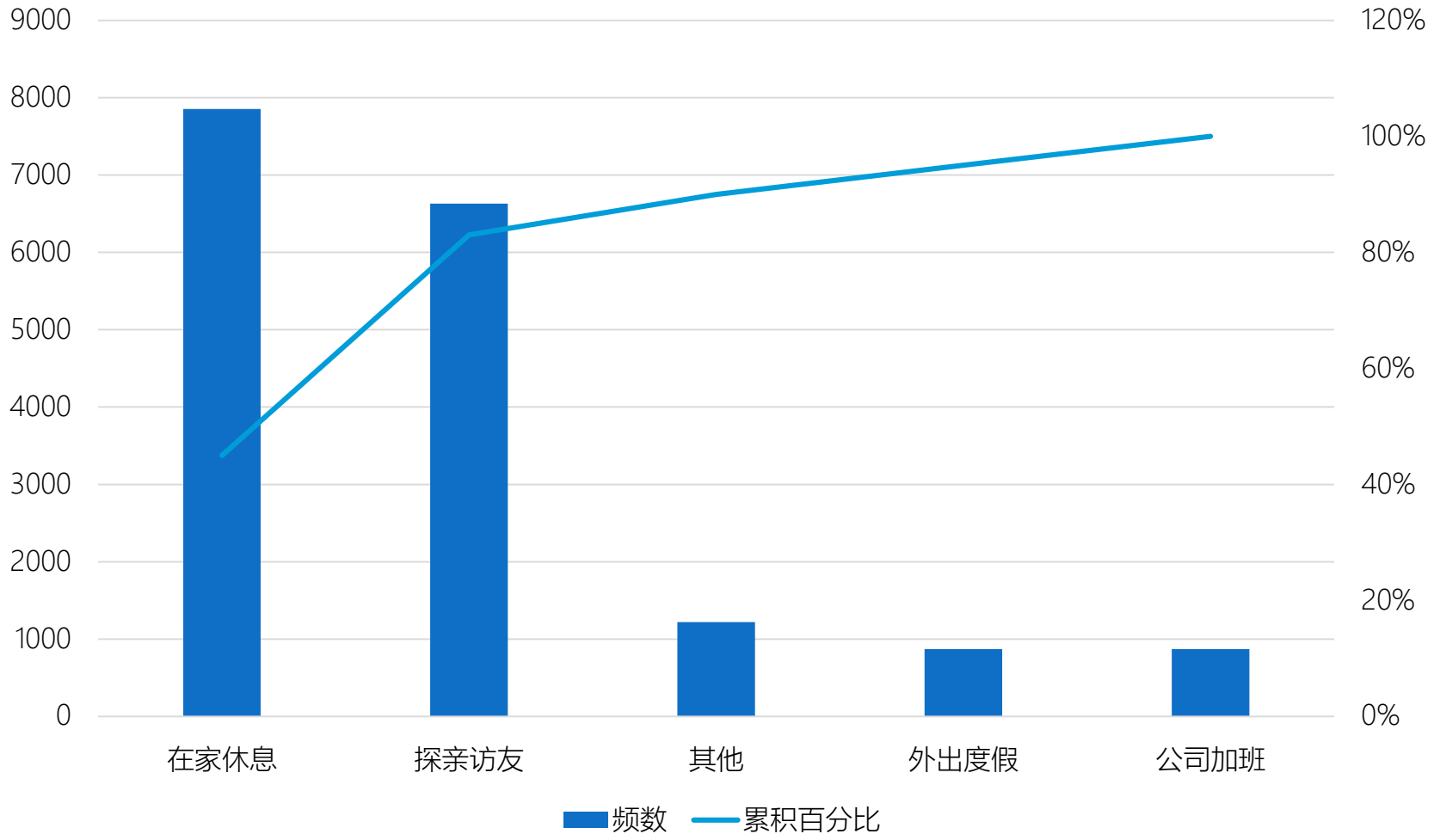


折线图

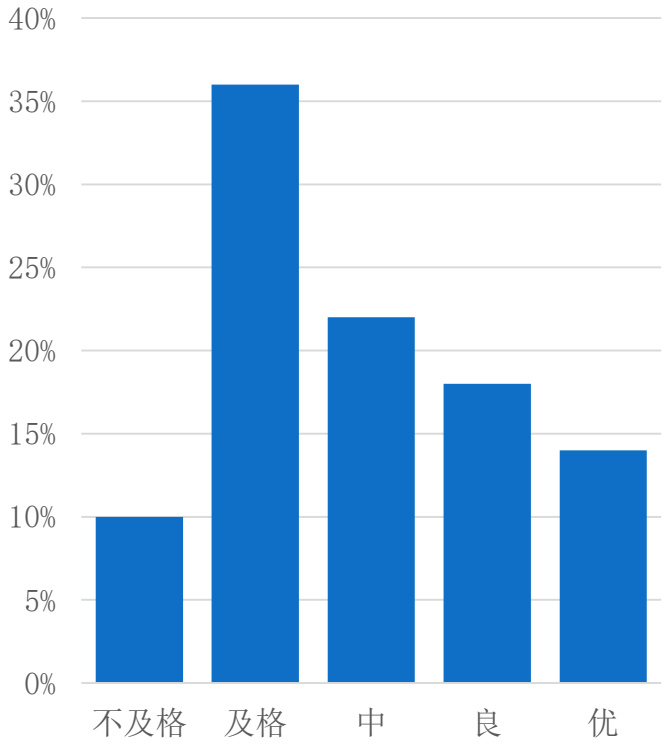

50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
在家休息
探亲访友
外出度假
公司加班
其他

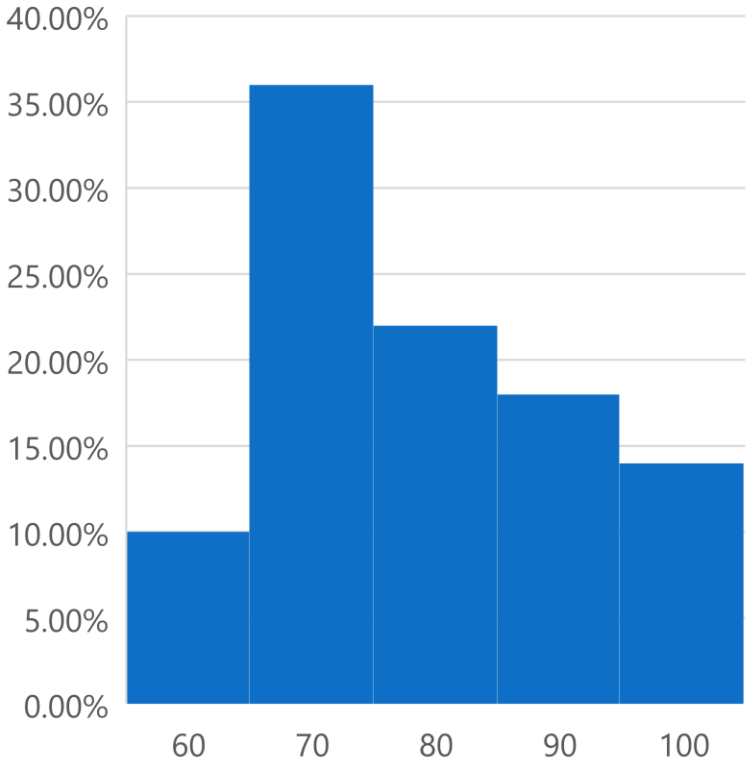

120%
100%
80%
60%
40%
20%
0%
在家休息
探亲访友
其他
外出度假
公司加班

帕累托图

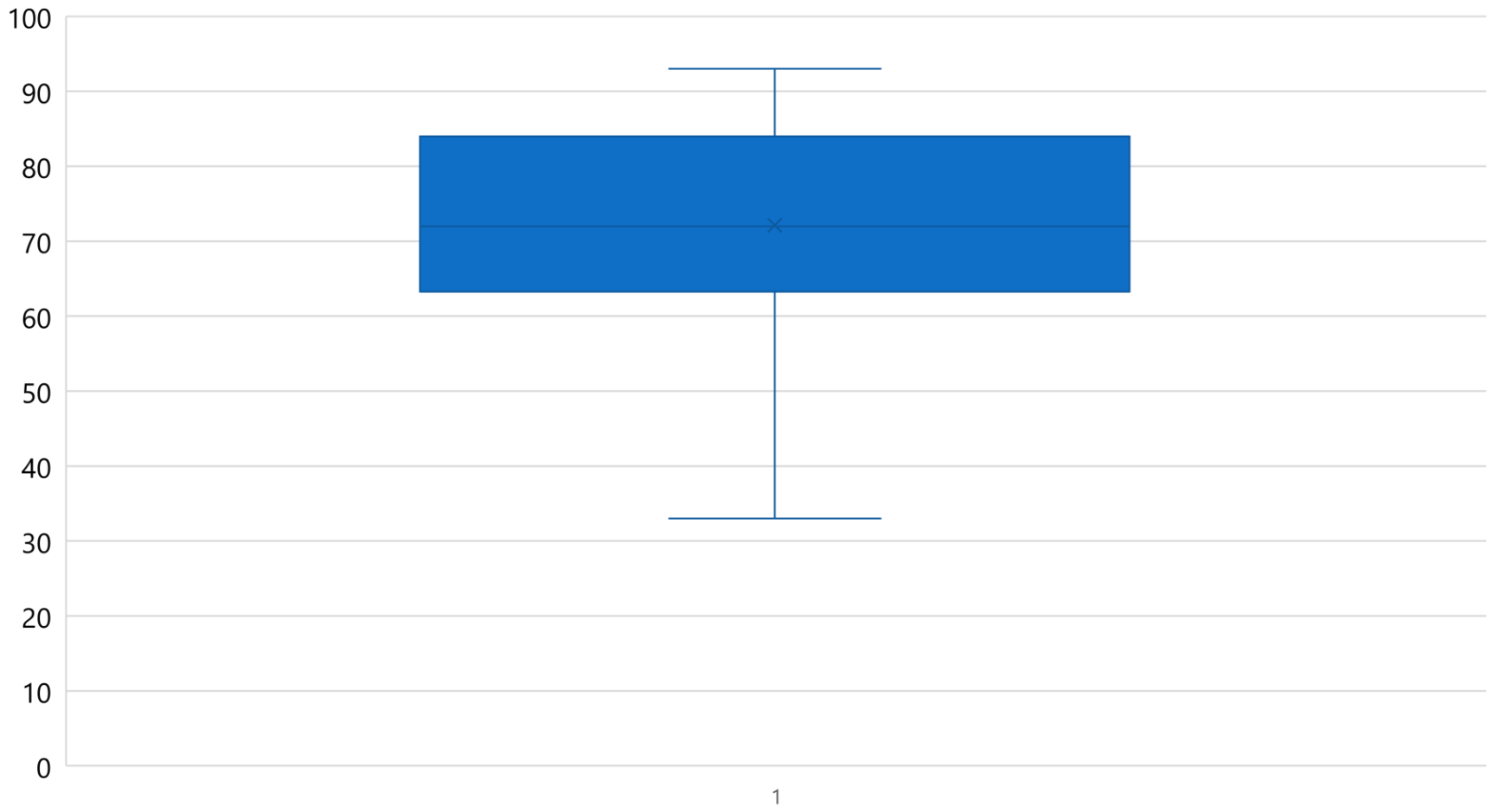

9000
8000
7000
6000
5000
4000
3000
2000
1000
0
120%
100%
80%
60%
40%
20%
0%
在家休息
探亲访友
其他
外出度假
公司加班
频数
累积百分比

柱状图与直方图对比

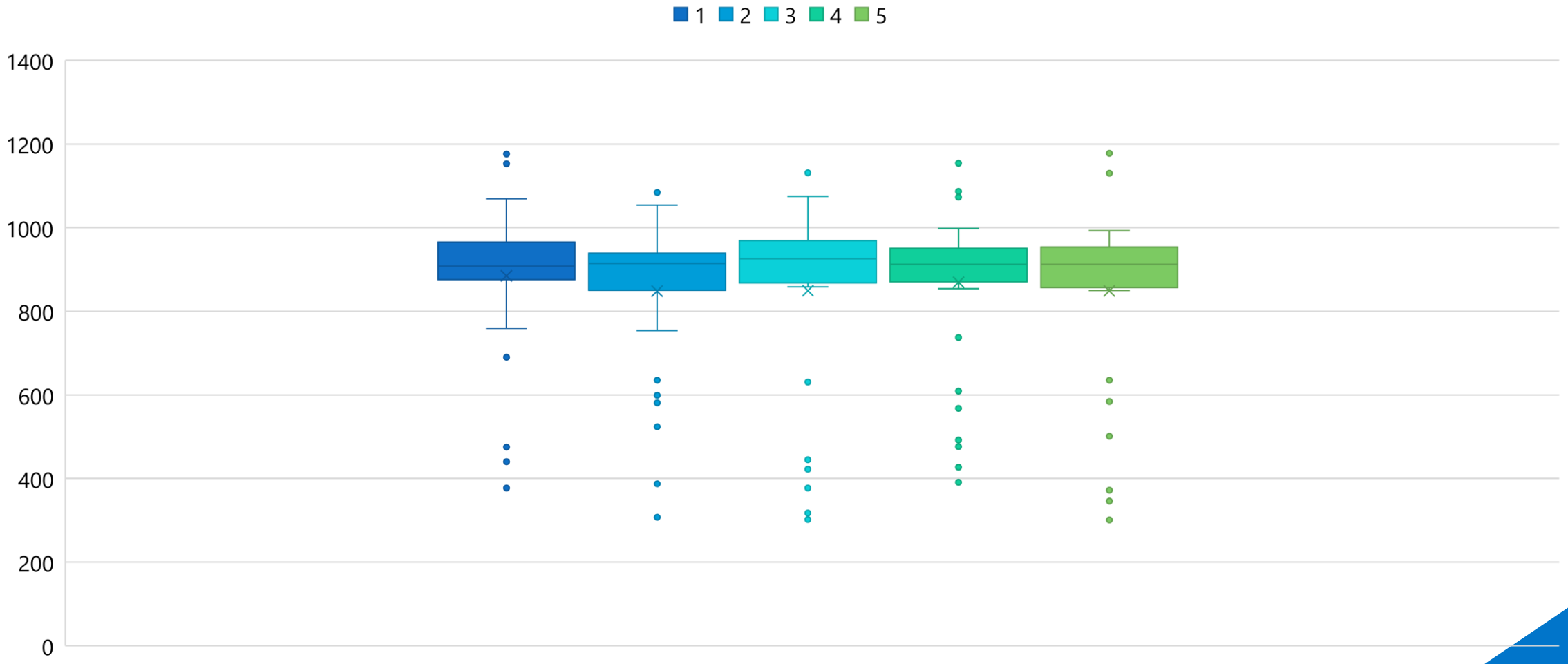

40%
35%
30%
25%
20%
15%
10%
5%
0%
不及格
及格
中
良
优


40.00%
35.00%
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
60
70
80
90
100

盒状图

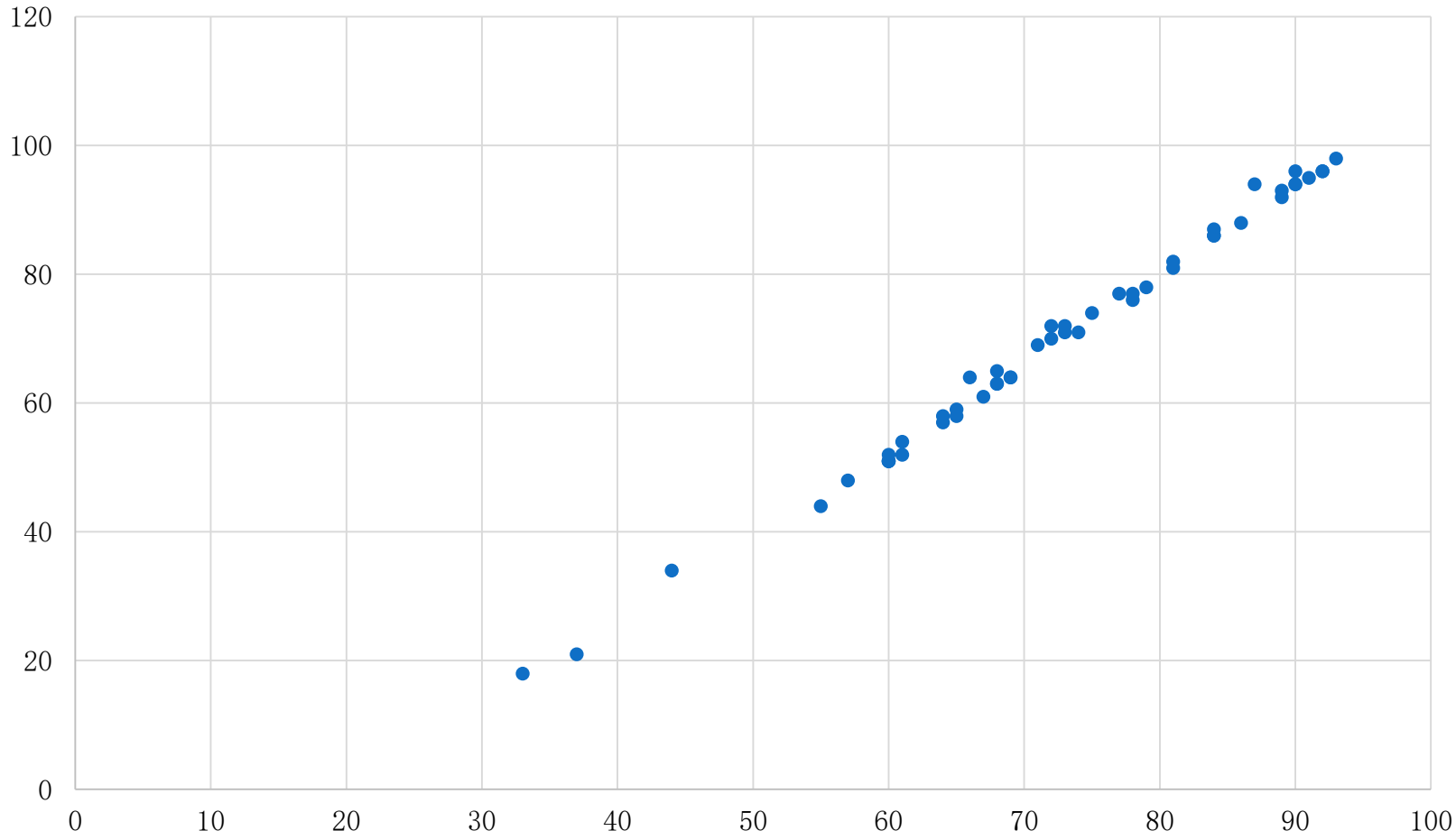


盒状图

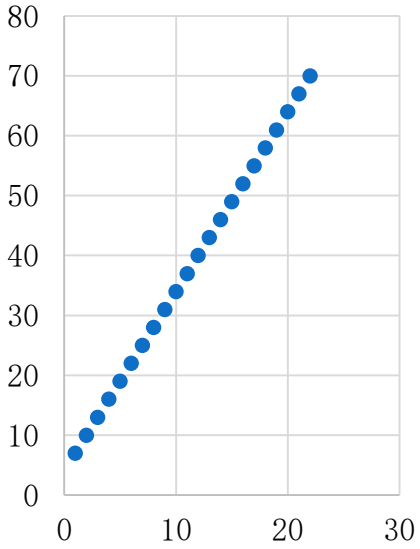


二维数据可视化

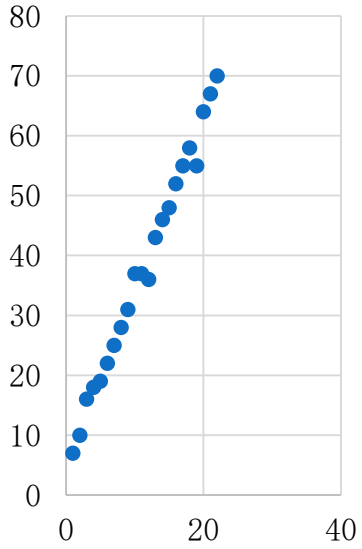
散点图



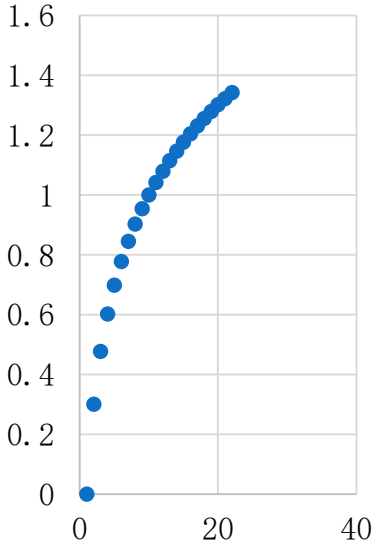
散点图



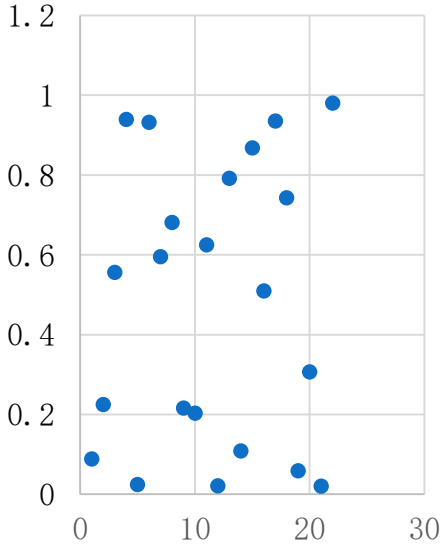
完全线性相关



线性相关



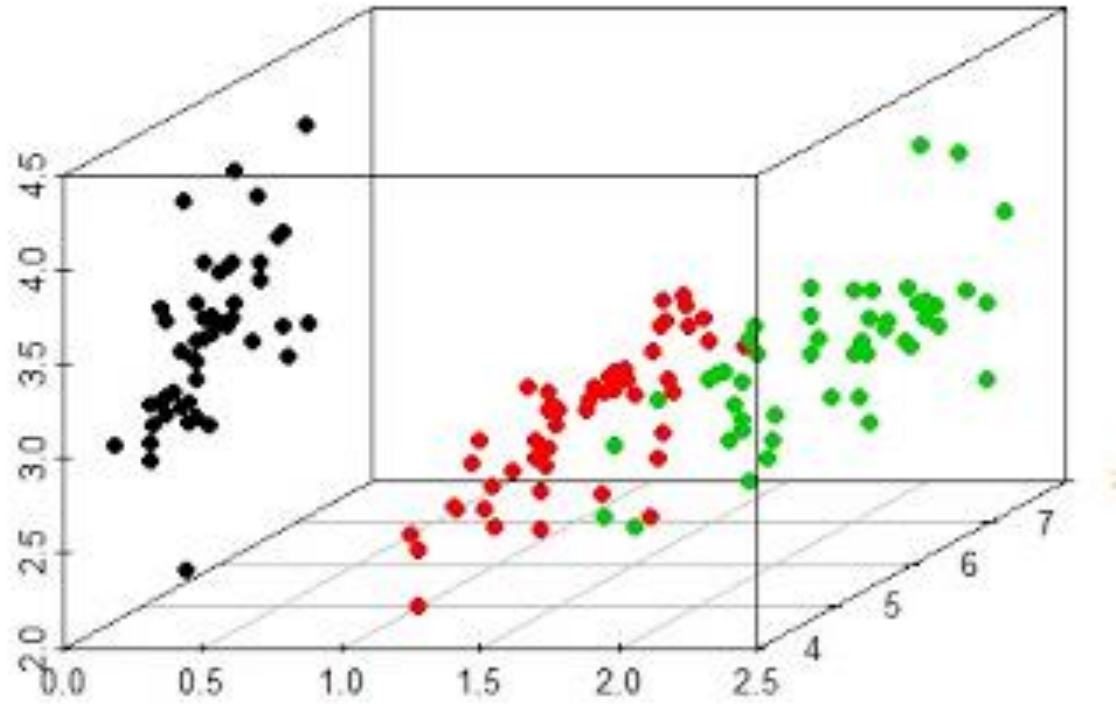
非线形相关



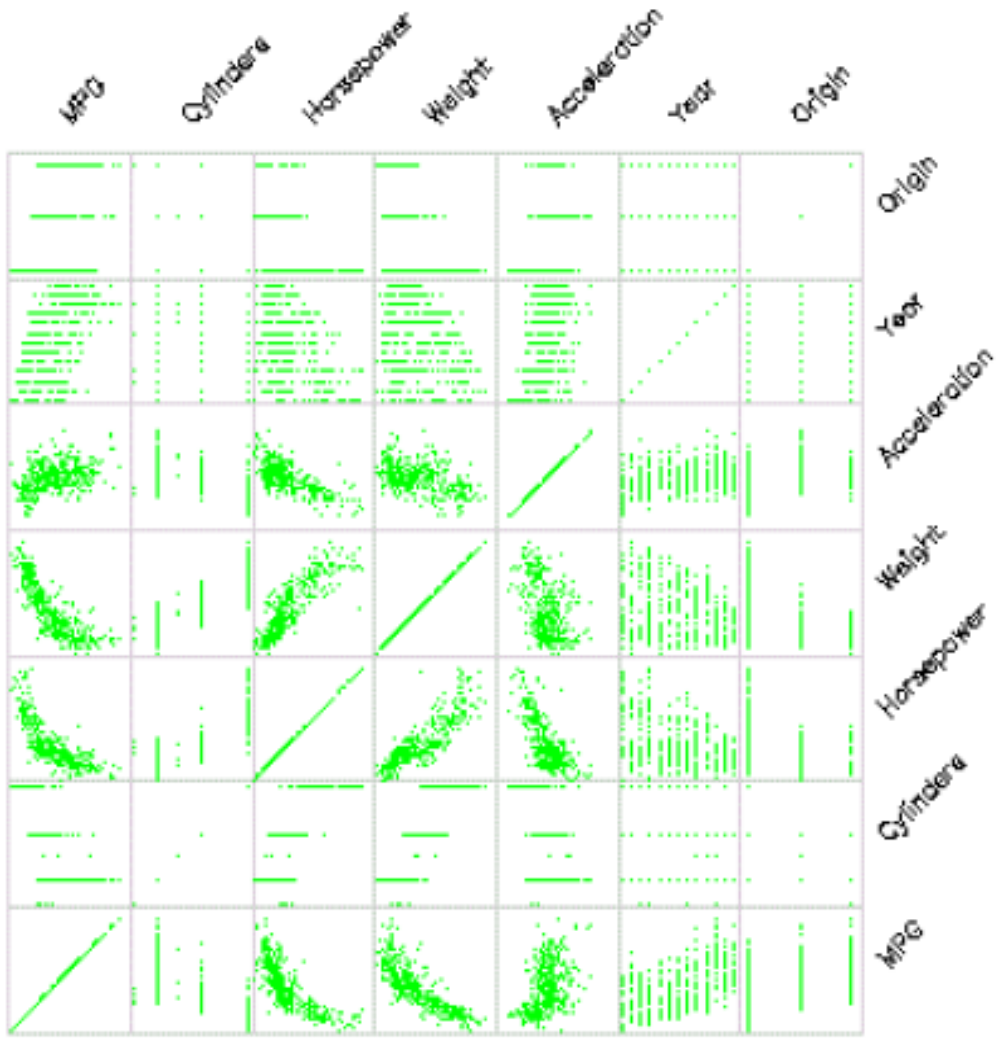
不相关

多维数据可视化

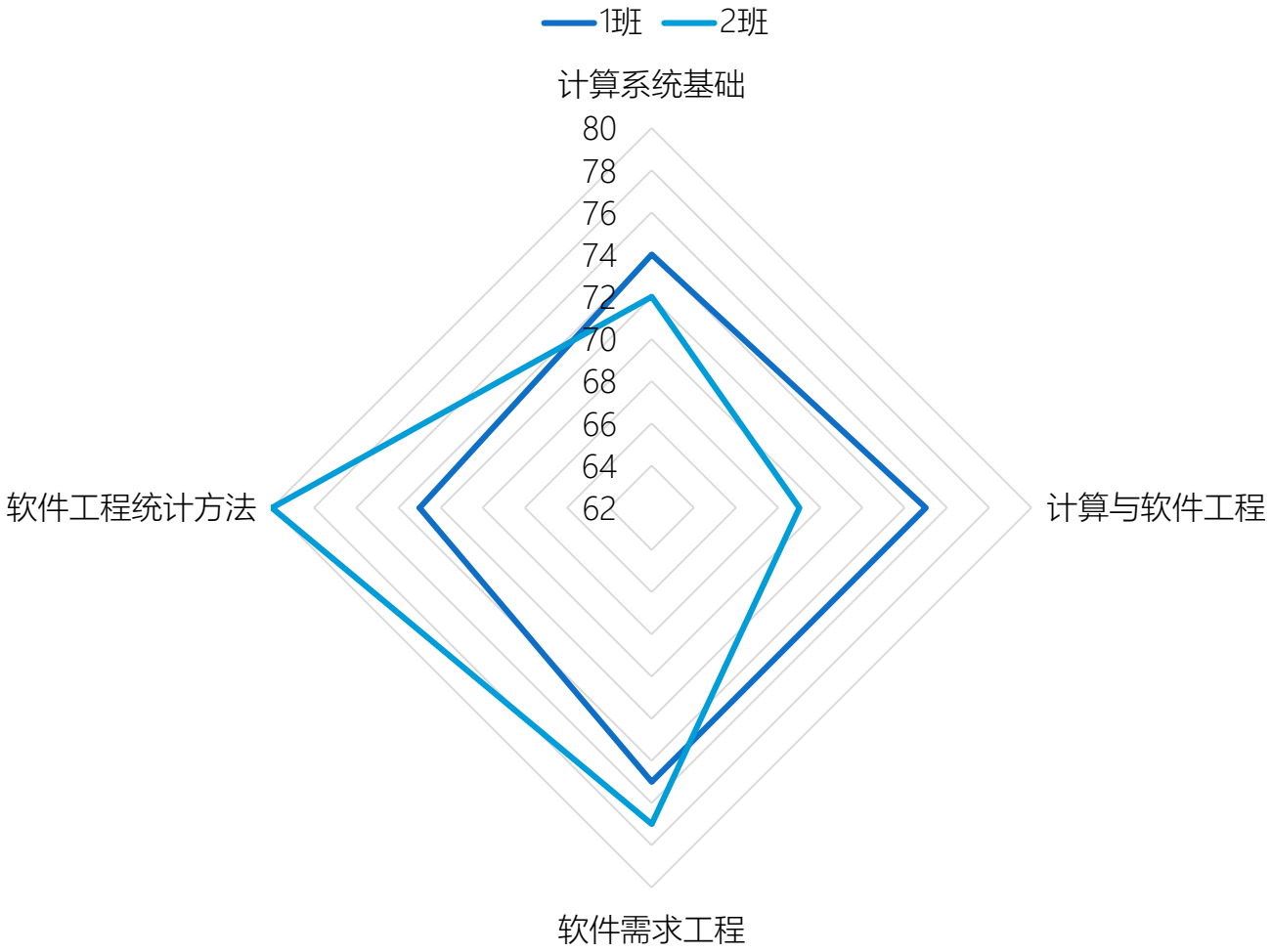
三维散点图



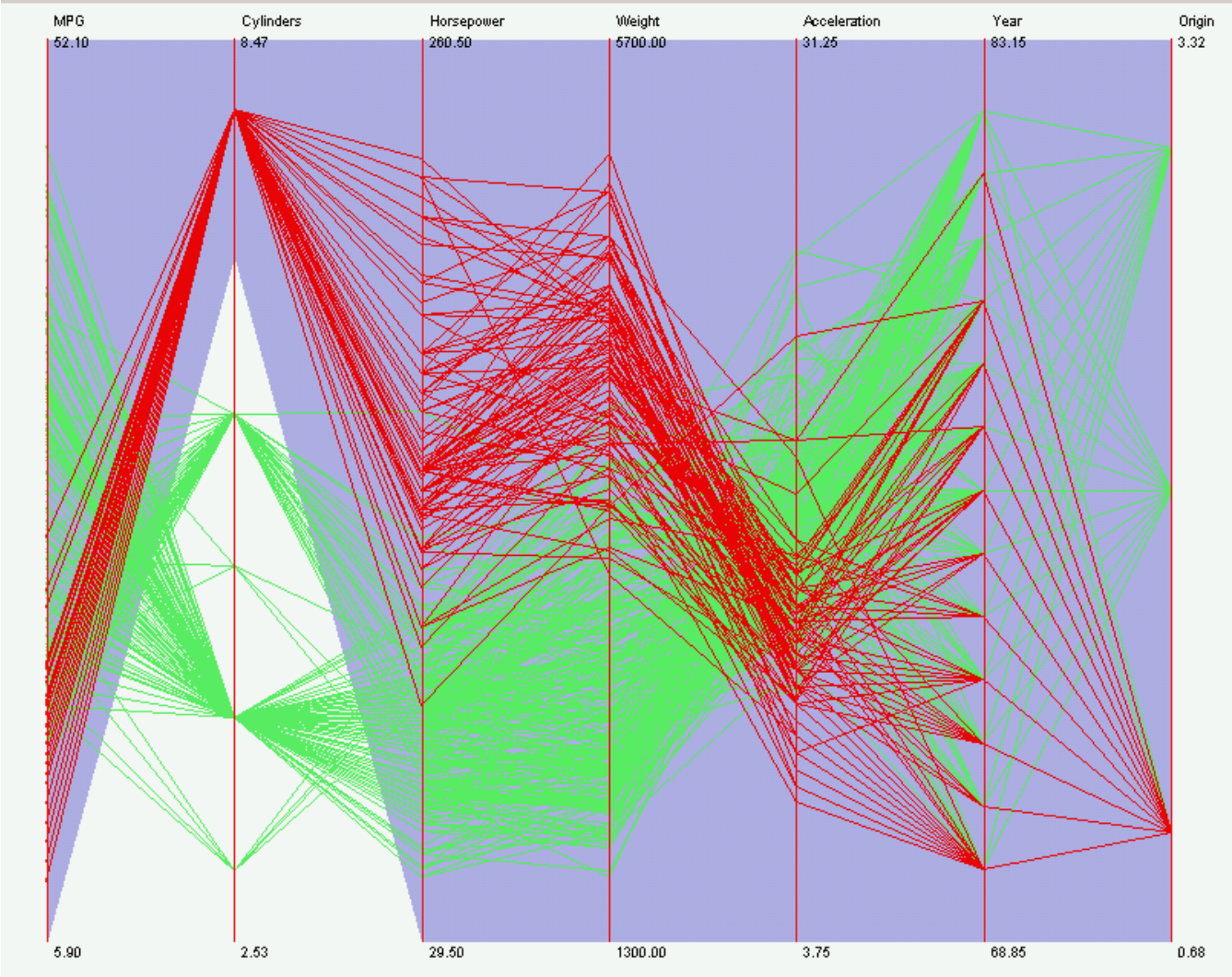
散点图矩阵


MPG
Cylinders
Horsepower
Weight
Acceleration
Year
Origin
Origin
Year
Acceleration
Weight
Horsepower
Cylinders
MPG

雷达图


1班
2班
计算系统基础
80
78
76
74
72
70
68
66
64
62
计算与软件工程
软件需求工程
软件工程统计方法

平行坐标

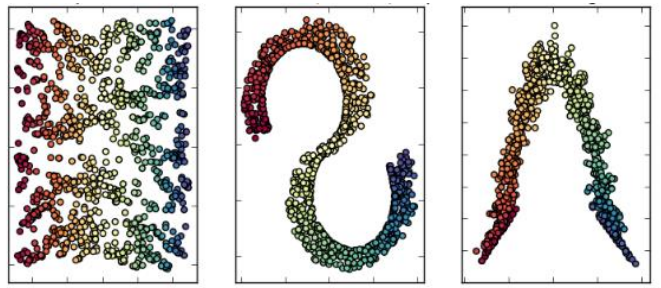

MPG
52.10
5.90
Cylinders
8.47
2.53
Horsepower
260.50
29.50
Weight
5700.00
1300.00
Acceleration
31.25
3.75
Year
83.15
68.85
Origin
3.32
0.68

高维数据可视化

高维数据可视化目的

- 数据探索：让自己理解数据
- 数据沟通：让他人更容易理解你

  从数据里面得出的观点/结论



高维数据可视化手段

尽可能在低维空间保持高维数据的结构特性