《用 Python 玩转数据》之用箱形图分析成绩数据

Dazhuang@NJU

“箱形图（英文：Box-plot），又称为盒须图、盒式图、盒状图或箱线图，是一种用作显示一组数据分散情况资料的统计图。因形状如箱子而得名。在各种领域也经常被使用，常见于品质管理。不过作法相对较繁琐。 箱形图于 1977 年由美国著名统计学家约翰·图基（John Tukey）发明。 它能显示出一组数据的最大值、最小值、中位数、下四分位数(25%)及上四分位数(75%)。”——来自维基百科

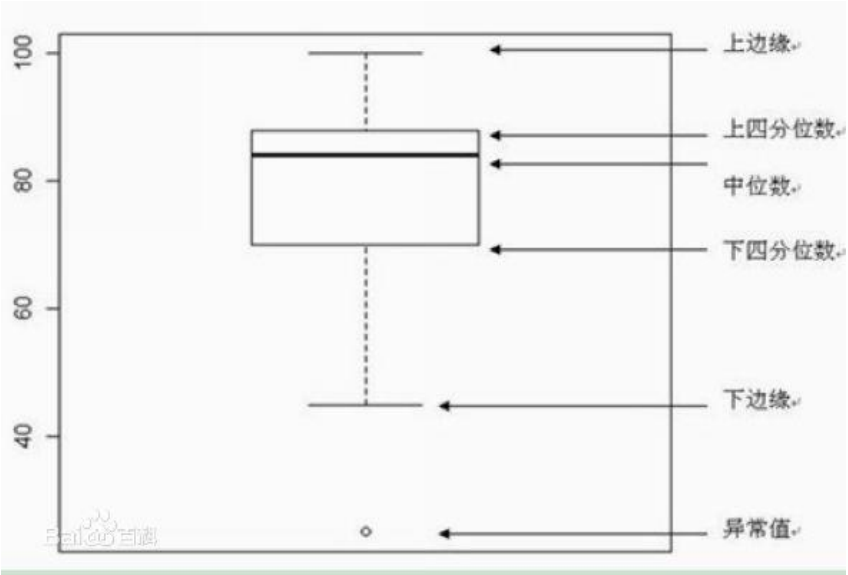


图 1 来自百度百科

把以下数据复制到 excel 工作表中（文件名为 score.xlsx）：

	Maths	English	Python	Music	Physics	Chemistry	PE
Wang	88	64	96	85	90	81	95
Ma	92	99	95	94	92	94	90
Liu	91	87	99	95	95	92	70
Qian	78	99	75	81	83	88	92
Meng	88	78	98	84	70	95	98
Song	100	95	100	92	98	95	65

利用 pandas 的 read_excel()方法读出数据保存到一个 DataFrame 中，并绘制相应的箱形图，观察生成的箱形图思考以下几个问题：

（1）哪些课程的成绩分布比较集中，哪些比较分散？你是通过什么来观察到的？

（2）哪些课程的成绩分布比较平均（对称）？你是通过什么来观察到的？

（3）哪些课程的成绩总体较好？哪些总体较差？你是通过什么来观察到的？

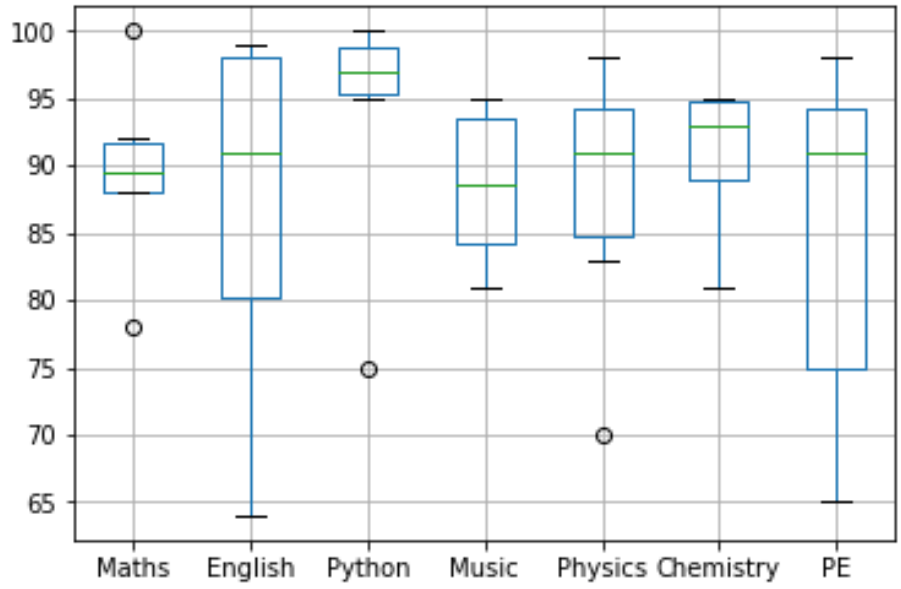
（4）哪些课程的成绩存在与总体情况不相符的值？你是通过什么来观察到的？

【参考程序和分析见下一页】

参考代码：

```
import pandas as pd
scores = pd.read_excel('scores.xlsx')
scores.boxplot()
```

生成的图：



问题分析：

（1）Maths、Python 和 Chemistry 分布比较集中，而 English 和 PE 的分布则比较分散，从箱子的长度可以获得这些信息；

（2）在 7 门课程中，Python 的成绩分布最为均匀，PE 的成绩分布最不均匀，从中位数到上四分位和下四分位的距离比较可以获得这些信息；

（3）Python 的总体情况最好，Maths、English 和 Music 的总体情况不太理想，从中位数的位置可以获得这些信息，分析数据可以发现这 3 门课程中各有 3 个同学没有达到 90 分，PE 虽然从图中看平均值应该不高，但其总体情况较好，分析数据可以发现有 2 个同学考了低分但 90 以上的有 4 个同学；

（4）Maths、Python 和 Physics 中分别有 2 个、1 个、1 个异常值（离群点），箱形图中会将这些异常值单独列出，一是为了一目了然地表明数据中的异常，二是为了不因为这些少数的异常数据导致整体特征的偏移。