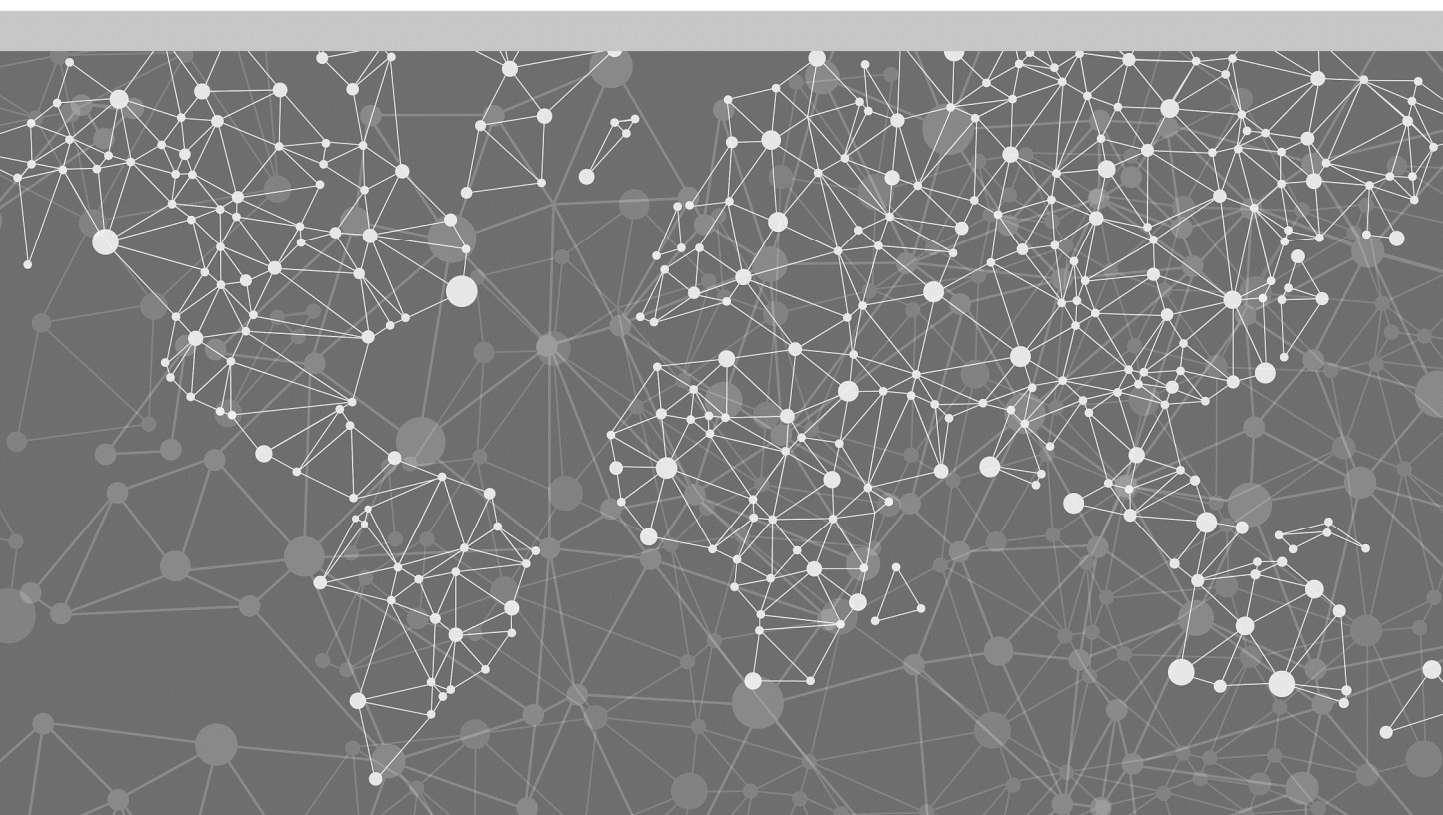
第一部分

数据分析基础知识



第 1 章　数据收集与分析软件

1.1　数据收集过程

1.1.1　数据的类型

数据是采用某种计量尺度对事物进行计量的结果，采用不同的计量尺度会得到不同类型的数据。通常按数据的收集途径可将数据进行如下分类：

1.1.1.1　按度量尺度分

(1)定性数据(也称计数数据，qualitative data)

定性数据是对度量事物进行分类的结果。数据表现为类别，用文字来表述，如性别、区域、产品分类等。假如某班学生按性别分为男、女两类，那么性别就构成了一个定性变量。

性别：女，男，男，女，男，男，女，男，女，男，…，女，男，女，女，男，男，女，男，女

具体见 1.1.2 节例 1.1。

(2)定量数据(也称计量数据，quantitative data)

定量数据是对度量事物的精确测度。结果表现为具体的数值，如身高、体重、家庭收入、成绩等。假如测量某班每个学生的身高，这样身高就构成了一个定量变量。

身高：167，171，175，169，154，183，169，166，165，173，…，164，169，166，175，166，159，169，165

具体见 1.1.2 节例 1.1。

这类数据的详细分析参见王斌会编著的《数据统计分析及 R 语言编程》(第二版)。

1.1.1.2　按时间状况分

(1)横截面数据(也称截面数据，cross-section data)

横截面数据是指对变量在某一时点上收集的数据的集合，反映在相同或近似相同的时间点上收集的数据描述现象在某一时刻的变化情况。比如，2014 年我国各地区的国内生产总值、从业人员等数据：

地区	北京	天津	河北	山西	…	甘肃	青海	宁夏	新疆
生产总值	162.519	113.073	245.158	112.376	…	50.204	16.704	21.022	66.101
从业人员	1069.70	763.16	3962.42	1738.90	…	1500.30	309.18	339.60	953.34

当收集的数据有多个指标时，就形成了多元统计分析的数据格式，具体见 1.1.2 节例 1.2。

这类数据的详细分析参见王斌会编著的《多元统计分析及 R 语言建模》（第四版）。

(2)时间序列数据（也称动态数列，time series data）

时间序列数据是按照一定的时间间隔对某一变量在不同时间的取值进行观测得到的一组数据，反映在不同时间上收集到的数据描述现象随时间变化的情况。比如，收集 2015 年 6 月 3 日至 2018 年 5 月 31 日的沪深 300 指数的收盘价数据，这些数据就是一个时间序列数据：

日期 2015-6-3 2015-6-4 2015-6-5 2015-6-8 ... 2018-5-28 2018-5-29 2018-5-30 2018-5-31

收盘价 5143.590 5181.416 5230.552 5353.751 ... 3833.26 3804.01 3723.37 3802.38

具体见 1.1.2 节的例 1.3。

这类数据的详细分析参见王斌会编著的《计量经济学模型及 R 语言应用》一书。

1.1.2 数据的收集

数据收集有一定的格式，当对一个观察指标测量了每一观察单位的数据时，通常以向量的形式展现，x：$x_1, x_2, \cdots, x_n$。

当对每一观察单位测量了多个指标时，通常以双向表的矩阵形式展现，即

$$\boldsymbol{X}：\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_m$$

这里 $\boldsymbol{X}_j(j=1, 2, \cdots, m)$ 为 $n\times1$ 向量，$\boldsymbol{X}=(x_{ij})_{n\times m}$，如下所示。

$$\begin{array}{c} \\ 1 \\ 2 \\ \cdots \\ n \end{array}\begin{array}{c} \begin{array}{cccc} X_1 & X_2 & \cdots & X_m \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{11} & x_{22} & \cdots & x_{21} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \end{array}$$

不同领域对该数据的观察单位和指标的叫法不同：数学上称它们为行(row)和列(column)的二维数组或矩阵，统计学上称它们为观测(observation)和变量(variable)的数据集，数据库中称它们为记录(record)和字段(field)的数据表，人工智能中称它们为示例(example)和属性(attribute)的数据集。

为了使大家将注意力集中在如何进行数据分析，而不是将精力花在对数据的收集和输入上，本书采用一种新的数据分析策略，即通篇使用几组数据讲解如何进行数据分析。

1.1.2.1 单变量数据收集

这类数据通常都是一个个单独的数据变量，都可单独拿来进行数据分析。

【例 1.1　调查数据】

为了解某高校 52 名研究生的一些基本情况和对开设数据分析课程的一些看法，共收集了这些学生的八项指标(有时为了方便编程运算，也可将变量名改成英文或拼音形式)：

学生编号(定性变量，按年份、学院、专业、序号排列，简记为【学号】，也可记为 id)。

学生性别(定性变量，简记为【性别】，也可记为 sex)。

学生身高(定量变量，单位 cm，简记为【身高】，也可记为 height)。

学生体重(定量变量，单位 kg，简记为【体重】，也可记为 weight)。

学生个人年消费支出额(定量变量，单位千元，简记为【支出】，也可记为 outcome)。

开设课程的必要性(定性变量，简记为【开设】，也可记为 setup)。

是否学过相关课程(定性变量，简记为【课程】，也可记为 course)。

是否学过或用过何种数据分析软件(定性变量，简记为【软件】，也可记为 software)。

数据由变量及其观测值所组成。本例共有 8 个变量：学号、性别、身高、体重、支出、开设、课程、软件。

表 1-1 是 52 名研究生的个人和开课信息调查数据，按照该数据格式，每行为一个观测单位(样品)，每列为一个指标(变量)。于是就构成了表 1-1 的数据集，该数据保存在 PyDm_data.xlsx 文档的基本数据表单【BSdata】中。

表 1-1　52 名研究生的开课信息调查数据

学号	性别	身高	体重	支出	开设	课程	软件
1510248008	女	167	71	46.0	不清楚	都未学过	No
1510229019	男	171	68	10.4	有必要	概率统计	Matlab
1512108019	女	175	73	21.0	有必要	统计方法	SPSS
1512332010	男	169	74	4.9	有必要	编程技术	Excel
1512331015	男	154	55	25.9	有必要	都学习过	Python
1516248014	男	183	76	85.6	不必要	编程技术	Excel
1516352030	女	169	71	9.1	有必要	编程技术	Excel
1516171019	女	166	66	2.5	不必要	都未学过	Excel
1516391008	女	165	69	35.6	不必要	都未学过	Excel
1520395019	男	173	63	22.8	有必要	统计方法	R
1520100029	男	184	82	10.3	有必要	都学习过	SAS
1520324035	男	163	66	13.0	有必要	概率统计	Matlab
1522186005	男	162	63	9.8	有必要	都学习过	SPSS
1522160006	女	168	72	35.3	不必要	统计方法	SPSS
1522274026	女	164	66	50.5	有必要	统计方法	SPSS
1523376027	男	180	81	64.1	有必要	统计方法	Excel
1523368030	女	158	63	20.6	不清楚	都学习过	Excel
1524225006	男	179	75	5.8	有必要	编程技术	Python
1524105026	女	163	65	69.4	有必要	编程技术	Python
1524286013	男	160	62	4.8	有必要	都未学过	R

续表

学号	性别	身高	体重	支出	开设	课程	软件
1525235027	女	168	70	8.2	有必要	都学习过	R
1525352033	男	185	83	5.1	有必要	都学习过	SPSS
1526177005	男	174	76	15.8	有必要	概率统计	Excel
1526196010	男	167	72	9.8	不清楚	统计方法	SPSS
1527173011	女	160	62	11.5	不必要	都学习过	Matlab
1527237032	女	163	65	19.4	有必要	统计方法	R
1527289024	男	155	50	10.8	有必要	概率统计	SPSS
1529107020	男	178	78	8.9	不清楚	概率统计	Matlab
1529314037	男	170	70	15.1	有必要	概率统计	SAS
1529245023	男	164	58	21.9	有必要	统计方法	Excel
1529365032	男	172	71	10.4	有必要	都学习过	SPSS
1530273031	男	178	77	35.6	不必要	统计方法	R
1530243029	男	186	87	9.5	不必要	都未学过	No
1531364037	女	171	69	7.3	有必要	都学习过	Excel
1531316038	女	156	56	52.8	有必要	统计方法	Excel
1532304031	女	166	68	47.9	不清楚	统计方法	SAS
1532208040	男	176	78	75.5	不必要	概率统计	Excel
1532292012	男	178	78	28.4	不必要	概率统计	No
1532185004	女	155	54	13.4	不清楚	编程技术	Excel
1533219013	女	163	62	11.1	不清楚	概率统计	Matlab
1533384028	男	158	60	6.1	有必要	编程技术	R
1533172017	女	167	68	27.2	不必要	都未学过	Excel
1537288004	女	173	70	19.1	不清楚	编程技术	Python
1537359035	女	174	71	17.6	不清楚	概率统计	No
1438391022	女	164	62	10.3	有必要	编程技术	Python
1538399025	男	169	65	9.5	有必要	统计方法	SAS
1438120022	男	166	70	35.6	有必要	统计方法	R
1538319004	男	175	68	44.4	不清楚	统计方法	SAS
1538254010	女	166	65	5.3	不清楚	编程技术	Python
1540294017	女	159	58	71.4	不清楚	都学习过	SPSS
1540365026	女	169	73	5.5	有必要	统计方法	Excel
1540388036	女	165	67	56.8	不必要	概率统计	SAS

1.1.2.2 多元数据收集

这类数据也称横截面数据，主要用来研究多个变量间的关系，包括综合分析、分类分析等。

【例 1.2 综合数据】

为了解我国各地区对外贸易国际竞争力的情况，我们从各省(市、自治区)的对外贸易能力、对外贸易经济效益、贸易资本竞争力等方面选取了 8 个对外贸易国际竞争力的基础指标。

- 地区国内生产总值(百亿元，简记为【生产总值】，也可记为 Y)
- 从业人员人数(万人，简记为【从业人员】，也可记为 X1)
- 全社会固定资产投资额(百亿元，简记为【固定资产】，也可记为 X2)
- 实际利用外资总额(百亿元，简记为【利用外资】，也可记为 X3)
- 进出口贸易总额(亿美元，简记为【进出口额】，也可记为 X4)
- 工业企业新产品出口额(亿元，简记为【新品出口】，也可记为 X5)
- 国际市场占有率(‰，简记为【市场占有】，也可记为 X6)
- 对外贸易依存度(%，简记为【对外依存】，也可记为 X7)

这些指标基本覆盖了各省外贸国际竞争力的各方面，能够较好地反映各省国际竞争力水平。具体数据如表 1-2 所示。

表 1-2　我国 30 个省、市、自治区 2011 年对外贸易数据

地区	生产总值	从业人员	固定资产	利用外资	进出口额	新品出口	市场占有	对外依存
北京	162.519	1069.70	55.789	196.906	3894.9	6470.51	2.635	1.55
天津	113.073	763.16	70.677	61.947	1033.9	7490.32	1.986	0.59
河北	245.158	3962.42	163.893	178.782	536.0	2288.19	1.276	0.14
山西	112.376	1738.90	70.731	104.945	147.6	1522.79	0.242	0.08
内蒙古	143.599	1249.30	103.652	54.426	119.4	342.36	0.209	0.05
辽宁	222.267	2364.90	177.263	155.296	959.6	4150.24	2.278	0.28
吉林	105.688	1337.80	74.417	58.843	220.5	746.94	0.223	0.13
黑龙江	125.820	1977.80	74.754	81.979	385.1	318.89	0.789	0.20
上海	191.957	1104.33	49.621	179.582	4373.1	10326.44	9.359	1.47
江苏	491.103	4758.23	266.926	261.118	5397.6	43928.94	13.953	0.71
浙江	323.189	3680.00	141.853	239.452	3094.0	25355.08	9.657	0.62
安徽	153.007	4120.90	124.557	92.613	313.4	2344.05	0.762	0.13
福建	175.602	2459.99	99.109	92.158	1435.6	7957.50	4.144	0.53
江西	117.028	2532.60	90.876	71.531	315.6	1301.04	0.977	0.17
山东	453.619	6485.60	267.497	223.057	2359.9	17688.02	5.614	0.34
河南	269.310	6198.00	177.690	147.022	326.4	2176.17	0.859	0.08
湖北	196.323	3672.00	125.573	113.434	335.2	1614.37	0.872	0.11
湖南	196.696	4005.03	118.809	106.234	190.0	1814.50	0.442	0.06
广东	532.103	5960.74	170.692	410.616	9134.8	56849.07	23.742	1.11
广西	117.209	2936.00	79.907	66.822	233.5	641.55	0.556	0.13
海南	25.227	459.22	16.572	18.885	127.6	185.49	0.113	0.33
重庆	100.114	1590.16	74.734	70.117	292.2	3928.45	0.886	0.19
四川	210.267	4785.50	142.222	162.007	477.8	1233.51	1.297	0.15
贵州	57.018	1792.80	42.359	39.441	48.8	308.65	0.134	0.06

续表

地区	生产总值	从业人员	固定资产	利用外资	进出口额	新品出口	市场占有	对外依存
云南	88.931	2857.24	61.910	66.849	160.5	257.76	0.423	0.12
陕西	125.123	2059.02	94.311	92.209	146.2	408.45	0.313	0.08
甘肃	50.204	1500.30	39.658	42.500	87.4	300.89	0.096	0.11
青海	16.704	309.18	14.356	10.488	9.2	0.30	0.030	0.04
宁夏	21.022	339.60	16.447	13.563	22.9	197.00	0.071	0.07
新疆	66.101	953.34	46.321	44.409	228.2	83.39	0.751	0.22

本书所选数据是中国 30 个省(市、自治区)(未包括西藏)2011 年的相关数据，数据来源于中国统计年鉴和各省统计年鉴，该数据存放在 PyDm_data.xlsx 文档的多元数据【MVdata】表单中。

1.1.2.3 时序数据的收集

时序数据是一类比较特殊的数据，也称为纵向数据，它对数据的格式有一定要求，特别是时间序列数据，须注意时间序列数据的输入格式。

【例 1.3 日期数据—股票数据】

今从某证券网站收集到 2015 年 6 月 3 日至 2018 年 5 月 31 日三年的沪深 300 指数的收盘价数据，如表 1-3 所示。这是一种典型的日期时间序列数据集，共 3 年 732 个数据，该数据存放在 PyDm_data.xlsx 文档的股票数据【TSdata】表中。

表 1-3 沪深 300 日收盘价数据

日 期	收 盘 价	日 期	收 盘 价	日 期	收 盘 价
2015-6-3	5143.590	2017-5-2	3426.58	…	…
2015-6-4	5181.416	2017-5-3	3413.13	2018-5-18	3903.06
2015-6-5	5230.552	2017-5-4	3404.39	2018-5-21	3921.24
2015-6-8	5353.751	2017-5-5	3382.55	2018-5-22	3906.21
2015-6-9	5317.461	2017-5-8	3358.81	2018-5-23	3854.58
2015-6-10	5309.112	2017-5-9	3352.53	2018-5-24	3827.22
2015-6-11	5306.590	2017-5-10	3337.70	2018-5-25	3816.50
2015-6-12	5335.115	2017-5-11	3356.65	2018-5-28	3833.26
2015-6-15	5221.167	2017-5-12	3385.38	2018-5-29	3804.01
2015-6-16	5064.820	2017-5-15	3399.19	2018-5-30	3723.37
…	…	2017-5-16	3428.65	2018-5-31	3802.38

进一步，我们还可以收集股票指数的时数据、分数据、秒数据、毫秒数据和微秒数据，这类数据就形成了高频数据，是一种大数据，限于篇幅，本文将不涉及。

上述的数据都是一些结构化数据，但随着大数据时代的来临，出现了大量的非结构化数据，这些数据的类型不只是由数字构成的数据库，还包括大量的文字、图像、影像和视频数据。

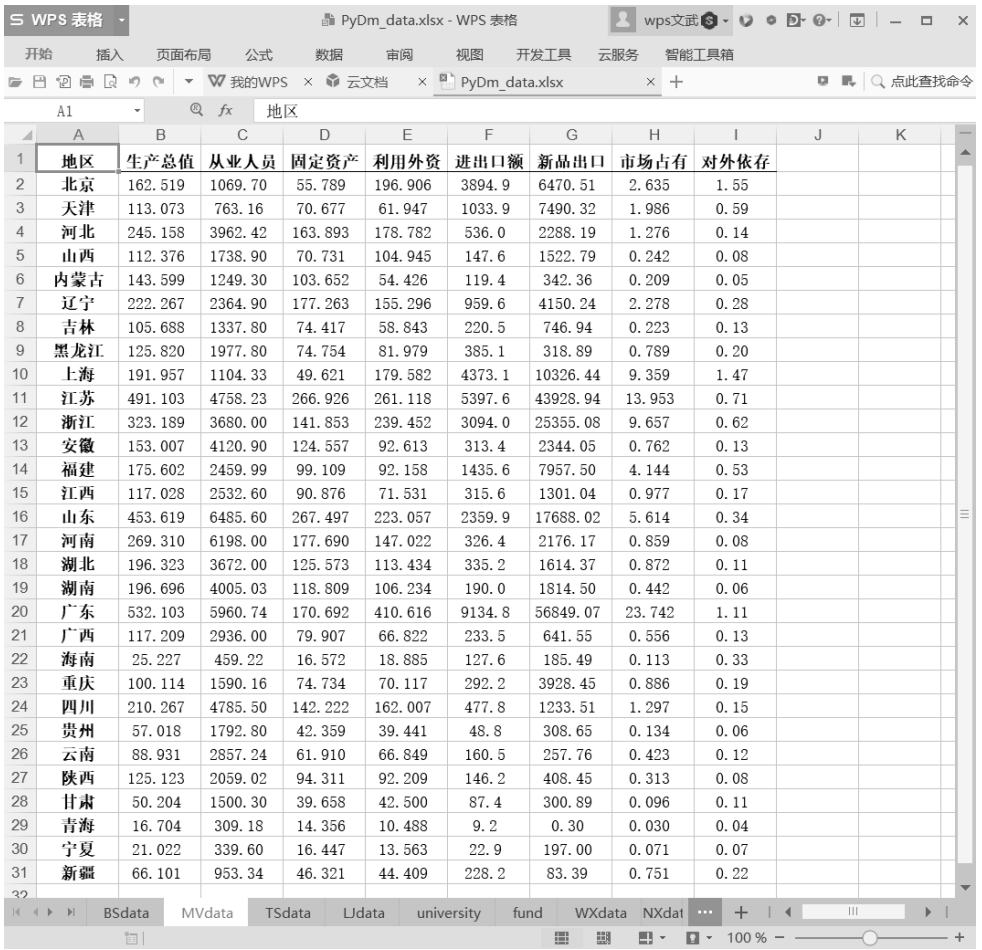
1.1.3 数据的管理

数据管理是利用计算机硬件和软件技术对数据进行有效的收集、存储、处理和应用的过程。对于一般的数据分析而言，电子表格软件已经足以胜任分析所需要的数据管理。最常用的电子表格软件有微软 Office 的 Excel 表格软件(收费)和金山 Office 的 WPS 表格软件(免费)。

1.1.3.1 电子表格管理数据

如果仅做一般数据管理，数据量不是特别大，而且要求系统免费、跨平台，那么首选的数据管理软件应该是 WPS 表格软件(WPS 表格是跟 Excel 兼容度最高的电子表格软件，但 WPS 是免费的，建议使用)。下面是采用 WPS 表格对上面数据的管理界面。

数据存放在 PyDm_data.xlsx 文档中，可登录 blog.leanote.com/PyDm 下载该数据。



地区	生产总值	从业人员	固定资产	利用外资	进出口额	新品出口	市场占有	对外依存
北京	162.519	1069.70	55.789	196.906	3894.9	6470.51	2.635	1.55
天津	113.073	763.16	70.677	61.947	1033.9	7490.32	1.986	0.59
河北	245.158	3962.42	163.893	178.782	536.0	2288.19	1.276	0.14
山西	112.376	1738.90	70.731	104.945	147.6	1522.79	0.242	0.08
内蒙古	143.599	1249.30	103.652	54.426	119.4	342.36	0.209	0.05
辽宁	222.267	2364.90	177.263	155.296	959.6	4150.24	2.278	0.28
吉林	105.688	1337.80	74.417	58.843	220.5	746.94	0.223	0.13
黑龙江	125.820	1977.80	74.754	81.979	385.1	318.89	0.789	0.20
上海	191.957	1104.33	49.621	179.582	4373.1	10326.44	9.359	1.47
江苏	491.103	4758.23	266.926	261.118	5397.6	43928.94	13.953	0.71
浙江	323.189	3680.00	141.853	239.452	3094.0	25355.08	9.657	0.62
安徽	153.007	4120.90	124.557	92.613	313.4	2344.05	0.762	0.13
福建	175.602	2459.99	99.109	92.158	1435.6	7957.50	4.144	0.53
江西	117.028	2532.60	90.876	71.531	315.6	1301.04	0.977	0.17
山东	453.619	6485.60	267.497	223.057	2359.9	17688.02	5.614	0.34
河南	269.310	6198.00	177.690	147.022	326.4	2176.17	0.859	0.08
湖北	196.323	3672.00	125.573	113.434	335.2	1614.37	0.872	0.11
湖南	196.696	4005.03	118.809	106.234	190.0	1814.50	0.442	0.06
广东	532.103	5960.74	170.692	410.616	9134.8	56849.07	23.742	1.11
广西	117.209	2936.00	79.907	66.822	233.5	641.55	0.556	0.13
海南	25.227	459.22	16.572	18.885	127.6	185.49	0.113	0.33
重庆	100.114	1590.16	74.734	70.117	292.2	3928.45	0.886	0.19
四川	210.267	4785.50	142.222	162.007	477.8	1233.51	1.297	0.15
贵州	57.018	1792.80	42.359	39.441	48.8	308.65	0.134	0.06
云南	88.931	2857.24	61.910	66.849	160.5	257.76	0.423	0.12
陕西	125.123	2059.02	94.311	92.209	146.2	408.45	0.313	0.08
甘肃	50.204	1500.30	39.658	42.500	87.4	300.89	0.096	0.11
青海	16.704	309.18	14.356	10.488	9.2	0.30	0.030	0.04
宁夏	21.022	339.60	16.447	13.563	22.9	197.00	0.071	0.07
新疆	66.101	953.34	46.321	44.409	228.2	83.39	0.751	0.22

1.1.3.2 数据库管理数据

当分析的数据量很大时，采用电子表格类软件有很大问题，须采用数据库来管理数据表格。

1.2 数据分析软件

1.2.1 数据分析软件简介

能做数据分析的软件有很多，如电子表格、SAS、SPSS、R、Python、Stata、Matlab、Eviews 等，下面简单介绍一下这些软件。

电子表格(Excel、WPS 等)不仅是数据管理软件，也是分析数据的入门工具。尽管其统计分析功能并不十分强大，但是它可以快速地做一些基本的数据分析工作，也可创建供大多数人使用的数据图表。由于电子表格在数据存量、图形样式、统计方法和统计建模方面功能受限，所以它们很难成为专业的数据分析软件。

SAS(Statistics Analysis System)是使用最为广泛的三大著名统计分析软件(SAS，SPSS 和 Splus)之一，被誉为统计分析的标准软件。SAS 是功能最为强大的统计软件，有完善的数据管理和统计分析功能，是熟悉统计学并擅长编程的专业人士的首选。

SPSS(Statistical Package for the Social Science)也是世界上著名的统计分析软件之一。SPSS 中文名为社会科学统计软件包，这是为了强调其社会科学应用的一面，而实际上它在社会科学和自然科学的各个领域都能发挥巨大作用。与 SAS 比较，SPSS 是非统计学专业人士的首选。

Matlab 是美国 MathWorks 公司出品的商业数学软件，是用于算法开发、数据可视化、数据分析及数值计算的高级技术计算语言和交互式环境，主要包括 Matlab 和 Simulink 两大部分。它在数值计算和模拟分析方面首屈一指，主要应用于工程计算、控制设计、信号处理与通信、图像处理、信号检测、金融建模设计与分析等领域。

Stata 是一套完整的、集成的统计分析软件包，可以满足数据分析、数据管理和统计图形的所有需要。Stata 12 增加了许多新的特征，比如结构方程模型(SEM)、ARFIMA、Contrasts、ROC 分析、自动内存管理等。Stata 适用于 Windows、Macintosh 和 Unix 平台计算机(包括 Linux)。Stata 的数据集、程序和其他的数据能够跨平台共享，且不需要转换，同样可以快速而方便地从其他统计软件包、电子表单和数据库中导入数据集。

Eviews 是美国 QMS 公司 1981 年发行的第 1 版 Micro TSP 的 Windows 版本，通常称为计量经济学软件包，是当今世界最流行的计量经济学软件之一。它可应用于科学计算中的数据分析与评估、财务分析、宏观经济分析与预测、模拟、销售预测和成本分析等。由于 Eviews 提供了一个很好的工作环境，能够迅速地进行编程、估计、使用新的工具和技术，所以它在计量经济建模方面有着广泛的应用。

从纯数据分析角度来说，应用最好的当属 S 语言的免费开源及跨平台系统 R 语言。R 语言是一个用于统计计算的很成熟的免费软件，也可以把它理解为一种统计计算语言，实际上很多人都直接称呼它为“R”，它比 C++，Fortran 等不知道简单了多少倍！如果你是一位数据分析的初学者，面对众多数据分析软件感到困惑且难以抉择，又想快速地掌

握统计计算、数据分析甚至目前比较流行的数据挖掘技术，那么首选的语言就是 R。

不过，R 语言对于初学编程和数据分析的人来说，入门还是有一定难度的，因为它还不是真正意义上的一般编程语言，所以现在流行“人生苦短，我用 Python”这样的说法，说明 Python 作为一种新兴的编程语言，已深入人心。现在我国许多地区高考试卷中都加入了 Python 编程的内容，一些中小学也开始开设 Python 编程课程。另外，由于 Python 博采众长，不断吸收其他数据分析软件的优点，并加入了大量的数据分析功能，它已成为仅次于 Java、C 及 C++的第四大语言，且在数据处理领域有超过 R 语言的趋势，所以本数据分析教程采用了 Python 作为分析工具。

综上所述，出于数据管理的方便，适用于一般数据分析的最好的数据管理软件应该是电子表格类软件(如微软 Office 的 Excel，金山 WPS 的表格等)，大量数据可以在一个工作簿中保存。所以，对于规模不是非常大的数据集，建议采用该方法来管理和编辑数据，而统计软件是我们进行数据分析不可或缺的工具。随着知识产权保护要求的不断提高，免费和开放源代码逐渐成为一种趋势，Python 正是在这个大背景下发展起来的，并逐渐成为数据分析的标准软件。考虑到微软的 Excel 必须购买正版，而 WPS 表格提供官方免费正版软件，笔者认为，通常的数据处理和分析工作用 WPS+Python 足矣。

1.2.2 Python 语言介绍

1.2.2.1 Python 简介

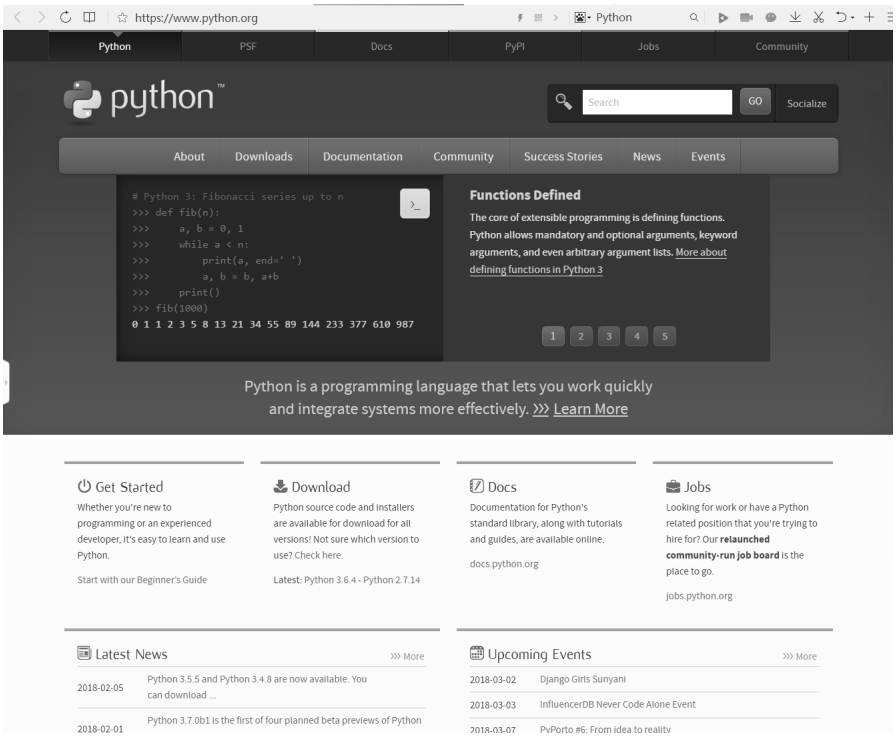
Python(英国发音：/'paɪθən/，美国发音：/'paɪθɑːn/)，是一种面向对象的解释型计算机程序设计语言，由荷兰人 Guido van Rossum 于 1989 年发明，第一个公开发行版发行于 1991 年。

Python 是纯粹的自由软件，源代码和解释器 CPython 遵循 GPL(GNU General Public License)协议。Python 语法简洁清晰，特色之一是强制用空白符(white space)作为语句缩进。

Python 具有丰富而强大的包，它常被昵称为“胶水语言”，能够把用其他语言制作的各种模块(尤其是 C/C++)轻松地联结在一起。常见的一种应用情形是，使用 Python 快速生成程序的原型(有时甚至是程序的最终界面)，然后对其中有特别要求的部分，用更合适的语言改写，比如，3D 游戏中的图形渲染模块性能要求特别高，就可以用 C/C++重写，然后封装为 Python 可以调用的扩展包。需要注意的是，在使用扩展包时可能需要考虑平台问题，某些扩展包可能不提供跨平台的实现。

由于 Python 语言的简洁性、易读性及可扩展性，在国外用 Python 做科学计算的研究机构日益增多，一些知名大学已经采用 Python 来教授程序设计课程。例如，卡耐基梅隆大学的编程基础、麻省理工学院的计算机科学及编程导论就使用 Python 语言讲授。众多开源的科学计算软件包都提供了 Python 的调用接口，如著名的计算机视觉包 OpenCV、三维可视化包 VTK、医学图像处理包 ITK。而 Python 专用的科学计算扩展包就更多了，如以下三个十分经典的科学计算扩展包：numpy、scipy 和 Matplotlib，它们分别为 Python 提供了快速数组处理、数值运算及绘图功能。因此，Python 语言及其众多的扩展包所构

成的开发环境十分适合工程技术、科研人员处理实验数据、制作图表，甚至开发科学计算应用程序。Python 的官方网站为 https://www.python.org/，在该网站可以下载 Python 软件和许多程序包，以及有关 Python 的资料。



1.2.2.2 Python 的特色

Python 是一种高层次的结合了解释性、编译性、互动性和面向对象的脚本语言，其设计具有很强的可读性。

① Python 是解释型语言：这意味着开发过程中没有了编译这个环节。

② Python 是交互式语言：这意味着可以在一个 Python 提示符下直接互动执行写程序。

③ Python 是面向对象语言：这意味着 Python 支持面向对象的风格或代码封装在对象中的编程技术。

④ Python 是初学者的语言：Python 对初级程序员而言，是一种友好易学的语言，它支持广泛的应用程序开发——从简单的文字处理到 WWW 浏览器再到游戏。

具体而言，Python 有如下一些特点。

① 简单、易学。

② 免费、开源。

③ 高层语言：封装内存管理等。

④ 可移植性：程序如果不使用依赖于系统的特性，那么无须修改就可以在任何平台上运行。

⑤ 解释性：直接从源代码运行程序，不再需要担心如何编译程序，使得程序更加易于移植。

⑥ 面向对象：支持面向过程的编程，也支持面向对象的编程。

⑦ 可扩展性：需要保密或者高效的代码，可以用 C 或 C++编写，然后在 Python 程序中使用。

⑧ 可嵌入性：可以把 Python 嵌入 C/C++程序，从而向程序用户提供脚本功能。

⑨ 丰富的包：包括正则表达式、文档生成、单元测试、线程、数据库、网页浏览器、CGI、FTP、电子邮件、XML、XML-RPC、HTML、WAV 文件、密码系统、GUI(图形用户界面)、Tk 和其他与系统有关的操作。

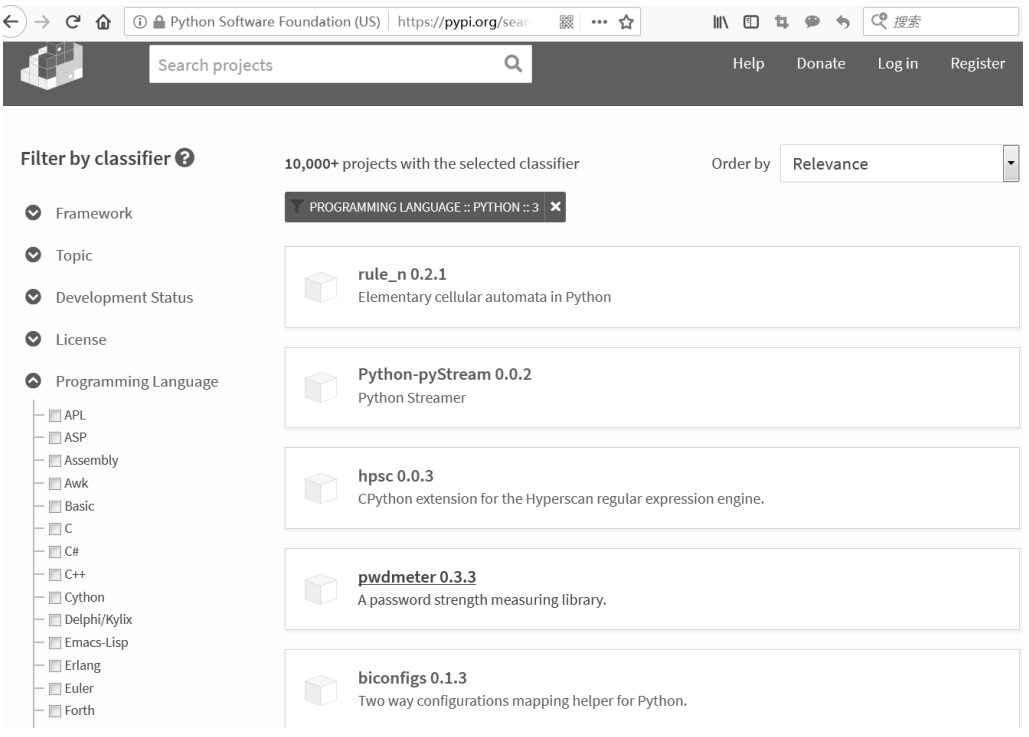
除标准包以外，还有许多其他高质量的包，如 wxPython、Twisted 和 Python 图像包等。

⑩ 概括性强：Python 确实是一种十分精彩又强大的语言，它合理地结合了高性能与使得编写程序简单有趣的特色。

⑪ 规范的代码：Python 采用强制缩进的方式，使得代码具有极佳的可读性。

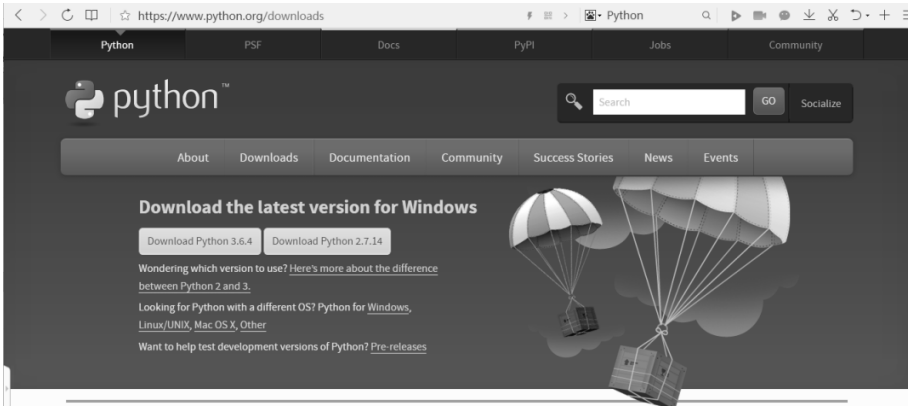
1.2.2.3 Python 的功能

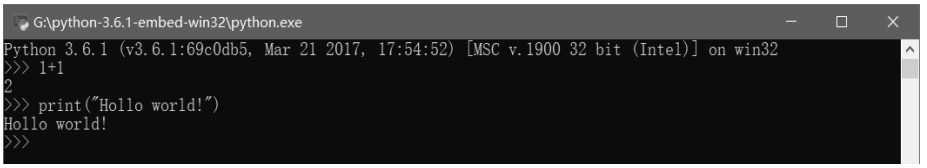
Python 最大也是其成为最流行的数据分析软件的特点就是，它包含大量的扩展包并拥有方便的二次开发功能。Python 的扩展包包罗万象，它所能完成的数据统计模型已经超出了任何其他商业统计软件。笔者做了一个统计，截至 2019 年 1 月，https://www.python.org/所列的扩展包达到 165797 个之多(包含几十万个数据分析方法)，除进行各种程序开发外，可完全满足进行数据分析之用。



1.2.2.4 Python 编程环境

Python 是一种强大的面向对象的编程语言，这样的编程环境需要使用者不仅熟悉各种命令的操作，还须熟悉 DOS 编程环境，而且所有命令执行完即进入新的界面，这给那些不具备编程经验或对统计方法掌握不够好的使用者造成了极大的困难。从 https://www.python.org/ 下载 Python 最新版，安装后只是一个不包括大量包的最基本的语言环境。本书采用基于 Anaconda 的 Jupyter 平台进行数据分析。

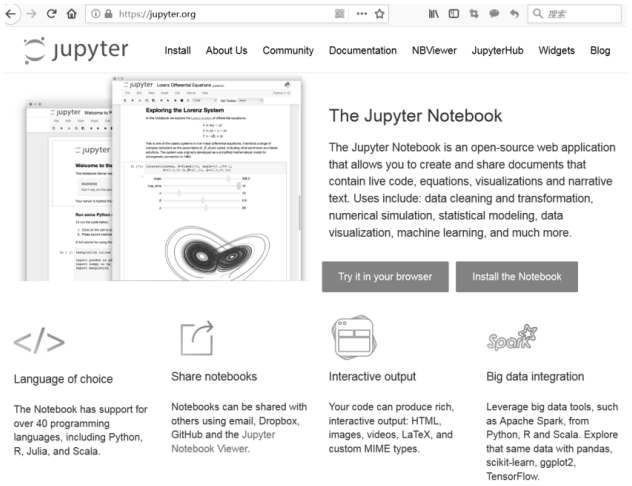




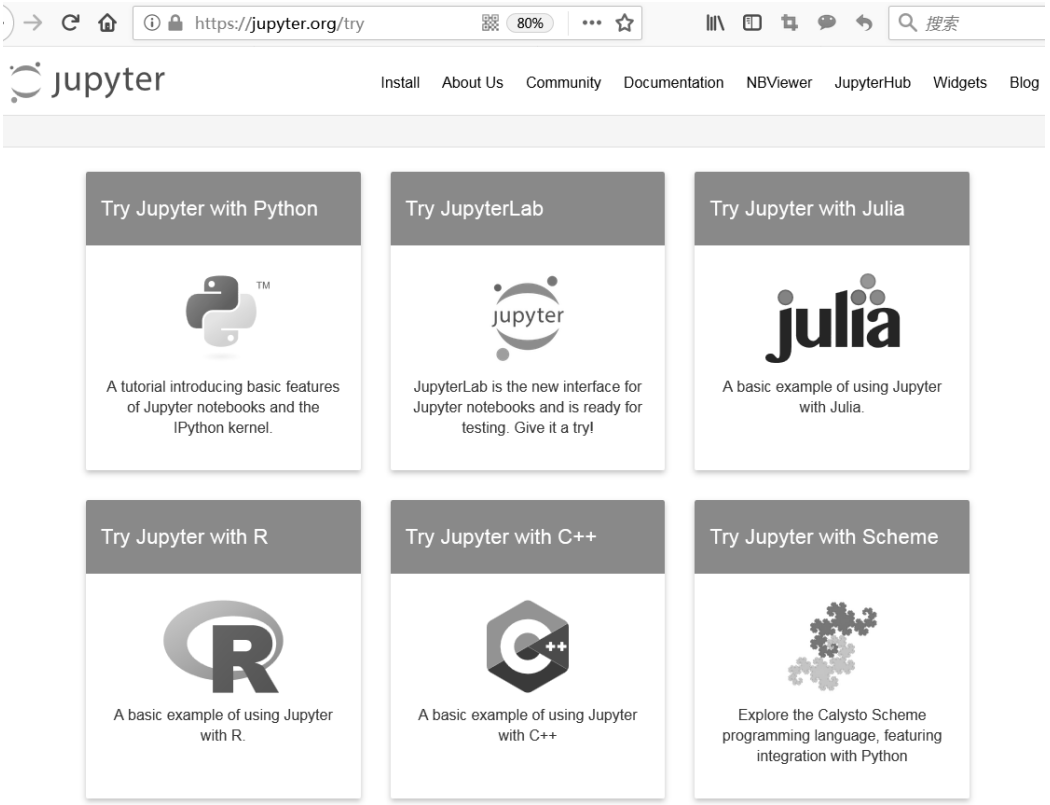
1.2.3 Python 在线平台

1.2.3.1 Jupyter 项目

随着网络技术的不断普及，建立基于大数据和云计算的 Web 应用平台势在必行。Jupyter 项目旨在开发跨几十种编程语言的开源软件、开放标准和用于交互式计算的服务。



Jupyter 项目目前提供了一个在线使用开源计算程序的云服务平台，可帮助大家快速使用 40 种以上编程语言，包括 Python、R、Julia 和 Scala 等，只要在网址中输入 https://jupyter.org/try 即可。



1.2.3.2 Jupyter Notebook

(1) Jupyter Notebook 简介

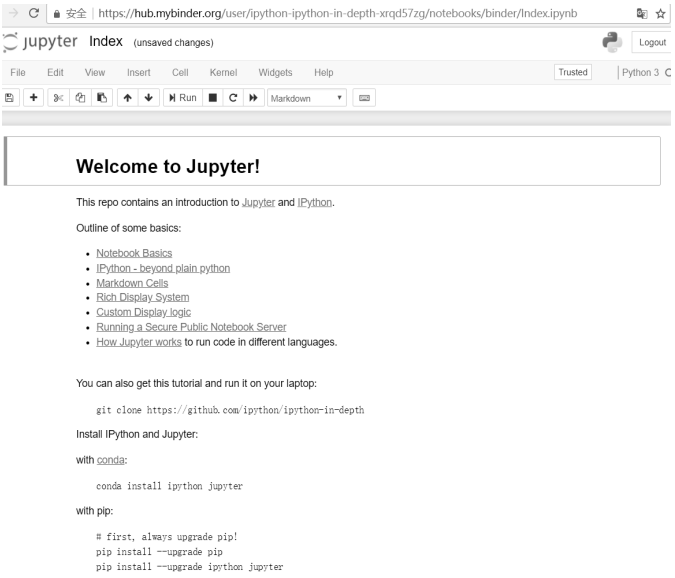
Jupyter Notebook 是一款开放源代码的 Web 应用程序，允许创建和共享包含实时代码、方程式、可视化和叙述文本的文档。用途包括数据清理和转换、数值模拟、统计建模、数据可视化、机器学习等。

使用 Jupyter Notebook，用户可以通过电子邮件、Dropbox、GitHub 和 Jupyter Notebook Viewer，将 Jupyter Notebook 分享给其他人。在 Jupyter Notebook 中，代码可以实时生成图像、视频、LaTeX 和 JavaScript。

数据挖掘领域的热门比赛 Kaggle 里的资料都是 Jupyter 格式的，本书也采用 Jupyter Notebook 格式。

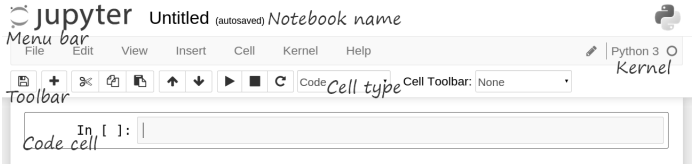
(2) Jupyter Notebook 的使用

Jupyter 社区提供浏览器版的 Jupyter Notebook，使用非常直观和方便，强烈推荐练习使用！但浏览器版只包含常用的程序包，一些复杂的程序包还得在本地安装版中使用。

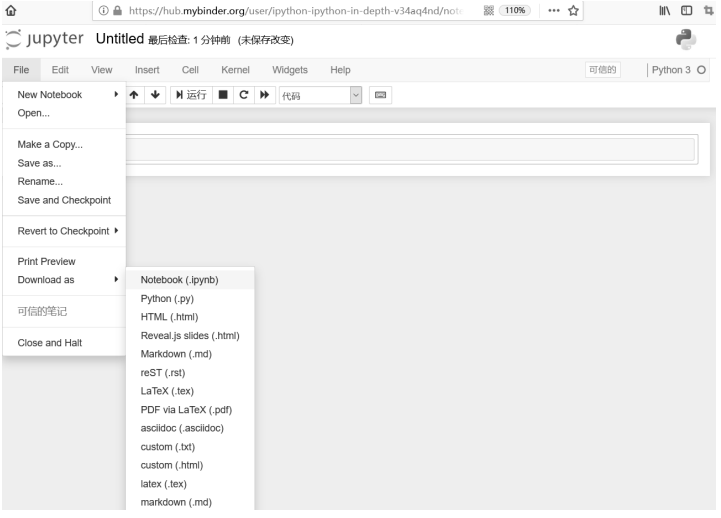


(3)新建 Jupyter Notebook 文档

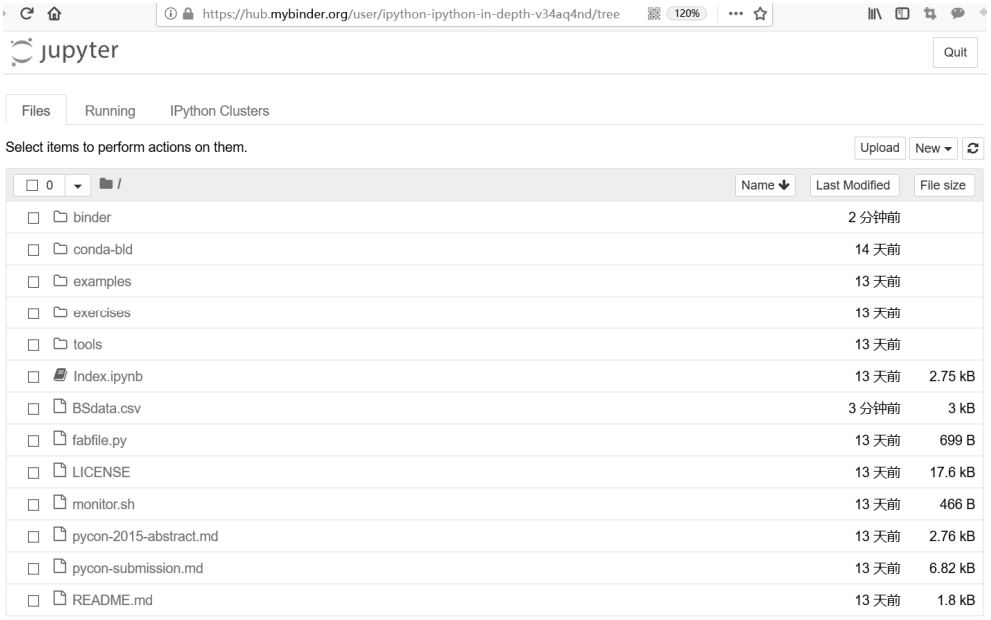
单击【New】按钮可建立相应的 Jupyter Notebook 文档语言文本，本书使用的是 Python3。建好文档后(这里默认文挡名为 Untitled.ipynb)就可以用 Python3 进行计算和分析了，也可以先建目录(Folder)，再建文档。



写文档时，cell 类型分成 markdown 和 code，可任意切换，直接写出；进行科学运算和画图时，numpy，scipy，pandas 等包以前都需要安装，现在全不用安装了。

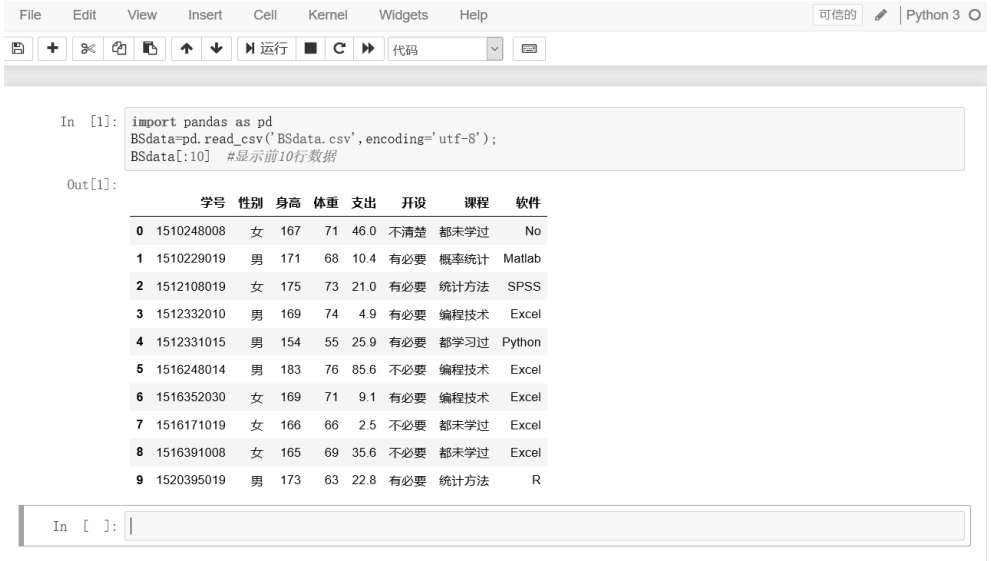


可以在文件管理菜单中修改(Rename...)之前新建的文档名，如将 Untitled.ipynb 修改为 myPython1.ipynb。也可以在文档菜单中下载并保存该文档，以备后用。



(4)上传文档与数据

由于 jupyter.org/try 是一个网络浏览器版，所以要使用自己的文档或数据，须事先上传【Upload】。比如，要用书中的基本数据进行分析，须上传 BSdata.csv 数据文档，然后就可以在 Jupyter Notebook 中使用了！



注意：对于文本数据，要留心数据的编码(encoding)格式！如果有中文名，要用'gb2312'或'utf-8'，但都得事先定义好！

(5) Jupyter Notebook 快捷键

Jupyter Notebook 有两种键盘输入模式。

① 编辑模式，允许往单元中输入代码或文本；这时的单元框线是绿色的。

② 命令模式，通过键盘输入运行程序命令；这时的单元框线是灰色的。

Shift+Enter：运行本单元，选中下一个单元；

Ctrl+Enter：运行本单元；

Alt+Enter：运行本单元，在其下插入新单元；

Y：单元转入代码状态；

M：单元转入 markdown 状态；

A：在上方插入新单元；

B：在下方插入新单元；

X：剪切选中的单元；

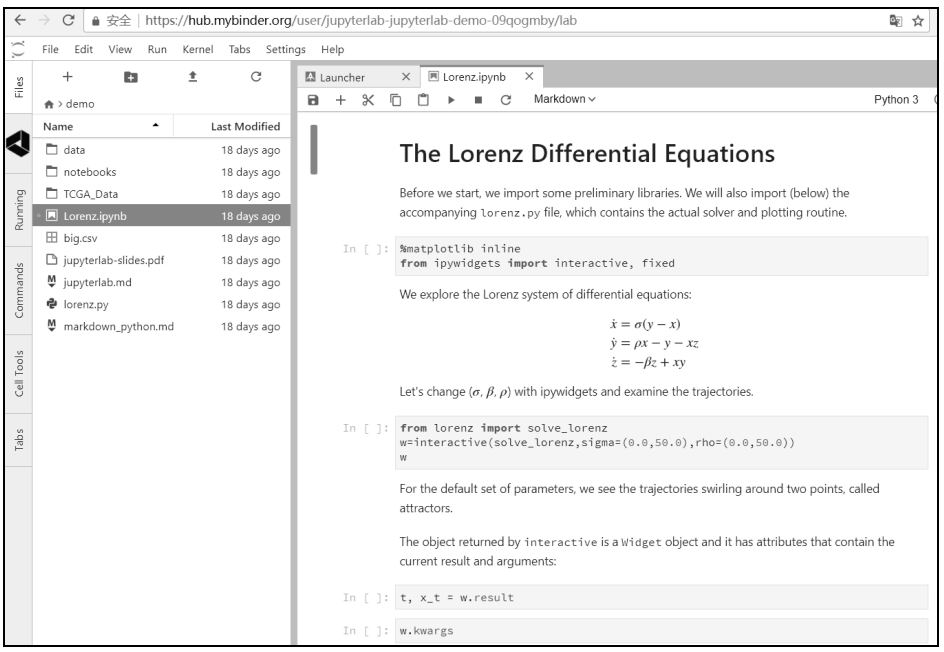
Shift +V：在上方粘贴单元。

这些快捷键也可在下面的 Jupyter Lab 中使用。

1.2.3.3 Jupyter Lab

Jupyter Lab 是一个名副其实的 IDE，且是一个基于网页的 IDE(保留了全部的 Notebook 特性)。

如果不想安装庞大的 Python 和 Jupyter Notebook，而只是想简单使用一下，那么可用 Jupyter 社区提供的浏览器版 Jupyter Lab，单击“试试 Jupyter Lab”即可使用，但限于网速，在线运行速度稍慢，建议使用本地安装版。



进入后的界面与平常使用的编程环境差别不大。

1.3 Python 编程基础

网上有大量的 Python 编程基础知识介绍，如

```
http://www.runoob.com/Python/Python-dictionary.html
```

请大家自行学习。由于本书重点介绍 Python 的数据分析，所以对 Python 编程的基础知识将不展开讨论。

1.3.1 Python 编程入门

1.3.1.1 Python 的工作目录

在使用 Python 时，一个重要设置是定义工作目录，即设置当前运行路径(全部数据和程序都将在该目录下工作)。例如，可以将 Python 工作目录设定为 D:\PyDm(先在 D 盘上建立目录 PyDm，然后在编程环境中使用)。

In [1]	#获得当前目录 %pwd #改变工作目录 %cd "D:\\PyDm" %pwd
Out [1]	"C:\user\1" D:\PyDm "D:\\PyDm"

1.3.1.2 Python 分析包

Python 具有丰富的数据分析模块，大多数做数据分析的人使用 Python 是因为其强大的数据分析功能。所有的 Python 函数和数据集是保存在包里面的。只有当一个包被安装并被载入(import)时，它的内容才可以被访问。这样做，一是为了高效(完整的列表会耗费大量的内存并且增加搜索的时间)；二是为了帮助包的开发者，防止命名和其他代码中的名称冲突。

由于 Anaconda 发行版已安装常用的数据分析包，所以我们只须调用即可。下面介绍几个 Python 常用的数据分析包，如表 1-4 所示。

表 1-4 Python 常用数据分析包

包 名	说 明	主要功能
math	基础数学包	提供函数，完成各种数学运算
random	随机数生成包	Python 中的 random 模块用于生成各种随机数
numpy	数值计算包	numpy(numeric python)是 Python 的一种开源的数值计算扩展，一个用 Python 实现的数值计算工具包。它提供许多高级的数值编程工具，如矩阵数据类型、矢量处理，以及精密的运算包。专为进行严格的数值处理而产生

续表

包　名	说　明	主要功能
scipy	数值分析包	提供很多科学计算工具包和算法，方便是易于使用，专为科学和工程设计的数值分析工具包。它包括统计、优化、整合、线性代数模块、傅里叶变换、信号和图像处理、常微分方程求解器等，包含常用的统计估计和检验方法
pandas	数据操作包	提供类似于 R 语言的 DataFrame 操作，非常方便。pandas 是面板数据(panel data)的简写。它是 Python 最强大的数据分析和探索工具，因金融数据分析工具而开发，支持类似 SQL 的数据增、删、改、查，支持时间序列分析，灵活处理缺失数据
statsmodels	统计模型包	statsmodels 可以补充 scipy.stats，是一个包含统计模型、统计测试和统计数据挖掘的 Python 模块。对每个模型都会生成一个对应的统计结果，对时间序列有完美的支持
matplotlib	基本绘图包	该包主要用于绘图和绘表，是一个强大的数据可视化工具，语法类似于 Matlab，是一个 Python 的图形框架，类似于 Matlab 和 R 语言。它是 Python 最著名的绘图库，提供了一整套和 Matlab 相似的命令 API，十分适合交互式制图。而且也可以方便地将它作为绘图控件，嵌入 GUI 应用程序中
sklearn	机器学习包	sklearn 是基于 Python 的机器学习工具模块，里面主要包含 6 大模块：分类、回归、聚类、降维、模型选择、预处理，如，使用 sklearn.decomposition 可进行主成分分解
beautifulSoup	网络爬虫包	beautifulsoup 是 Python 的一个包，最主要的功能是从网页抓取数据。BeautifulSoup 提供一些简单的、Python 式的函数，用来处理导航、搜索、修改分析树等功能。通过解析文档为用户提供需要抓取的数据，通过它可以很方便地提取出 HTML 或 XML 标签中的内容
networkx	复杂网络包	networkx 是一款 Python 的软件包，用于创造、操作复杂网络，以及学习复杂网络的结构、动力学及其功能。 通过它可以用标准或者不标准的数据格式加载或者存储网络，它可以产生许多种类的随机网络或经典网络，也可以分析网络结构、建立网络模型、设计新的网络算法、绘制网络等

注意：安装程序包和载入程序包是两个概念，安装程序包是指将需要的程序包安装到电脑中，载入包是指将程序包调入 Python 环境中。程序包的安装(通常在命令行状态：)>>>pip install pandas。

Python 调用包的命令是 import，如要调用上述包，可用

```
import math
import random
import numpy
import scipy
import pandas
import matplotlib
```

这些包中的函数，可直接使用包名加“.”。如要用 matplotlib 绘 plot 图，可用 matplotlib.plot(...)。

如要简化这些包的写法，可用 as 命令赋予别名，如

```
import numpy as np
import scipy as sp
```

```
import pandas as pd
import matplotlib as plt
```

这样 matplotlib.plot(...)可简化为 plt.plot(...)。

如要调用 Python 包中某个具体函数或方法，可使用 from ... import，例如，要调用 math 包中的开方、对数和 pi 函数，则用

```
from math import sqrt, log, pi
```

这样，可在程序中直接使用，如 sqrt(2)，等价于 math.sqrt(2)。

比如，要调用本书自定义函数文档 PyDm_fun.py 中的函数（见相关章节及附录），需按如下方式操作：

(1)安装自定义模块：将 PyDm_fun.py 文档复制到当前工作目录 D:\PyDm 下。

(2)加载自定义模块：import PyDm_fun as dm　# from PyDm_fun import *

(3)自定义函数调用：dm.mcor_test（X）# mcor_test（X）

In [2]	# 系统初始化	
	import numpy as np	#加载数值分析包
	np.set_printoptions(precision=4)	#设置 numpy 输出精度
	import pandas as pd	#加载数据操作包
	pd.set_option('display.width', 120)	#设置 pandas 输出宽度
	pd.set_option('display.precision',4)	#设置 pandas 输出精度
	import matplotlib.pyplot as plt	#加载基本绘图包
	plt.rcParams['font.sans-serif']=['KaiTi'];	#SimHei 黑体
	plt.rcParams['axes.unicode_minus']=False;	#正常显示图中负号
	import PyDm_fun as dm	#from PyDm_fun import * #加载自定义函数

1.3.1.3　Python 中的数据管理

目前，Python 最大的问题是数据管理，因为 Python 没有好用的数据管理器，其自带的数据管理器很不方便，所以，要用好 Python 软件，就得将 Python 与 Excel 等电子表格充分结合，发挥两者的优点，这样就可以事半功倍，这也是本书提出用“电子表格+Python”模式进行数据统计分析的原因。关于如何使用 Python 调用 Excel 等电子表格数据，参见 1.3.4.3 节。

1.3.2　Python 数据类型

1.3.2.1　Python 对象

Python 创建和控制的实体称为对象(object)，它们可以是变量、数组、字符串、函数或结构。由于 Python 是一种所见即所得的脚本语言，故不需要编译。在 Python 里，对象是通过名字创建和保存的。可以用 who 命令来查看当前打开的 Python 环境里的对象，用 del 删除这些对象。

(1)查看数据对象

In [3]	who
Out [3]	dm np pd