数据科学与大数据技术系列

Python 数据挖掘方法及应用

王斌会　王　术◎编著

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

数据科学与大数据技术系列

Python 数据挖掘方法及应用

王斌会　王　术　编著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

本书重点介绍 Python 语言在数据处理与数据挖掘方面的应用技巧，主要包括数据分析基础知识（数据收集与分析软件、数据挖掘的分析基础、简单数据的统计分析），数据分析高级方法（多元数据的综合分析、时序数据的模型分析），大数据基本处理方法（大数据分析基础应用、文献计量与科研评价、社会网络分析方法、数据分析编程平台）等内容。附录中还提供了 Python 数据分析相关方法和函数等，方便读者随时查看。本书内容丰富，图文并茂，可操作性强且便于查阅，主要面向数据分析的读者，能有效帮助读者提高数据处理与分析的水平，提升工作效率。书中的例子数据、习题数据及相关代码都可在作者的学习博客 http://blog.leanote.com/DaPy 下载使用，也可登录华信教育资源网 http://www.hxedu.com.cn 免费下载。

本书适合各层次的数据分析用户，既可作为初学者的入门指南，又可作为中高级用户的参考手册，同时也可作为各大中专院校和培训班的数据分析教材。

图书在版编目(CIP)数据

Python 数据挖掘方法及应用 / 王斌会，王术编著. —北京：电子工业出版社，2019.3
（数据科学与大数据技术系列）
ISBN 978-7-121-34495-4

Ⅰ. ①P… Ⅱ. ①王… ②王… Ⅲ. ①软件工具－程序设计－高等学校－教材 Ⅳ. ①TP311.561

中国版本图书馆 CIP 数据核字(2018)第 125853 号

策划编辑：秦淑灵
责任编辑：秦淑灵
印　　刷：北京季蜂印刷有限公司
装　　订：北京季蜂印刷有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编：100036
开　　本：787×1092　1/16　印张：13.5　字数：340 千字
版　　次：2019 年 3 月第 1 版
印　　次：2019 年 3 月第 1 次印刷
定　　价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010)88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：qinshl@phei.com.cn。

前　言

人类从农耕社会进入工业社会用了上千年时间，从工业社会进入信息社会用了一百多年时间，而从信息时代进入数据时代仅用了不到十年时间。随着互联网、物联网、云计算的不断深入应用，产生了大量的数据，这些数据的挖掘和分析应用，需要人们掌握数据分析技术。人类正全面进入大数据分析时代。

需要是发明之母。近年来，数据挖掘引起了信息产业界的极大关注，其主要原因是，存在大量的数据，可以被广泛使用，并且迫切需要将这些数据转换成有用的信息和知识。获取的信息和知识可以应用于各种领域，包括商务管理、生产控制、市场分析、工程设计和科学探索等。

“人生苦短，我要用 Python”，这是网上对 Python 评价最多的一句话，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。

本书重点介绍 Python 语言在数据处理与数据分析方面的应用技巧，涉及数据的整理、数据的输入和输出、探索性数据分析、基本数据分析、多元数据分析、时间序列数据分析、网络爬虫技术、社会网络分析、知识图谱和文献计量研究等数据分析方面的内容。附录中还提供了 Python 数据分析相关方法和函数等，方便读者随时查看。

全书分三部分，共 9 章内容。第一部分主要讲解数据分析基础知识，包括第 1、2、3 章，重点介绍数据收集与分析软件、数据挖掘的分析基础，以及简单数据的统计分析；第二部分讲解数据分析高级方法，包括第 4、5 章，主要介绍多元数据的综合分析和时序数据的模型分析；第三部分讲解大数据基本处理方法，包括第 6、7、8 章，重点介绍大数据分析基础应用、文献计量与科研评价、社会网络分析方法和数据分析编程平台。最后对 Python 的一些编程环境做了进一步介绍。

本书内容丰富，图文并茂，可操作性强且便于查阅，主要面向进行数据分析的读者，能有效地帮助读者提高数据处理与分析水平，提升工作效率。本书适合各层次的数据分析用户，既可作为初学者的入门指南，又可作为中高级用户的参考手册，同时也可作为各大中专院校和培训班的数据分析教材。

本书具有以下三大优点：

(1) 使用 Python 科学计算发行版 Anaconda，方便数据分析者使用。

读者可从 https://www.anaconda.com 下载安装并直接使用。

(2)公开本书自定义函数的源代码，使用者可以深入理解 Python 函数的编程技巧，用这些函数建立自己的开发包；并建立了本书的学习博客(http://blog.leanote.com/DaPy)，书中的例子数据、习题数据及相关代码都可直接在网上下载使用。

(3)采用网络化教学平台。Python 的基础版缺少一个面向一般人群的菜单界面，这对那些只想用其进行数据分析的使用者而言是一大困难，本书采用流行的 Python 网络分析平台 Jupyter(https://jupyter.org)，该平台可作为数据分析教学软件使用。

书中软件输出的坐标图多数没有标出横、纵坐标的量，目的是与软件界面保持一致。

本书在写作过程中得到了广东恒电信息科技股份有限公司的大力支持，该公司将为本书的实战操作提供可靠的实训环境支持，读者可以使用恒华大数据实训管理系统完成本书的实验操作。

本书由王斌会、王术共同完成，其中第 1～5 章由王斌会撰写，第 6～9 章由王术撰写，王斌会负责全书统稿。

由于作者知识和水平有限，书中难免有错误和不足之处，欢迎读者批评指正！

作　者

2019 年 1 月于暨南园

目　录

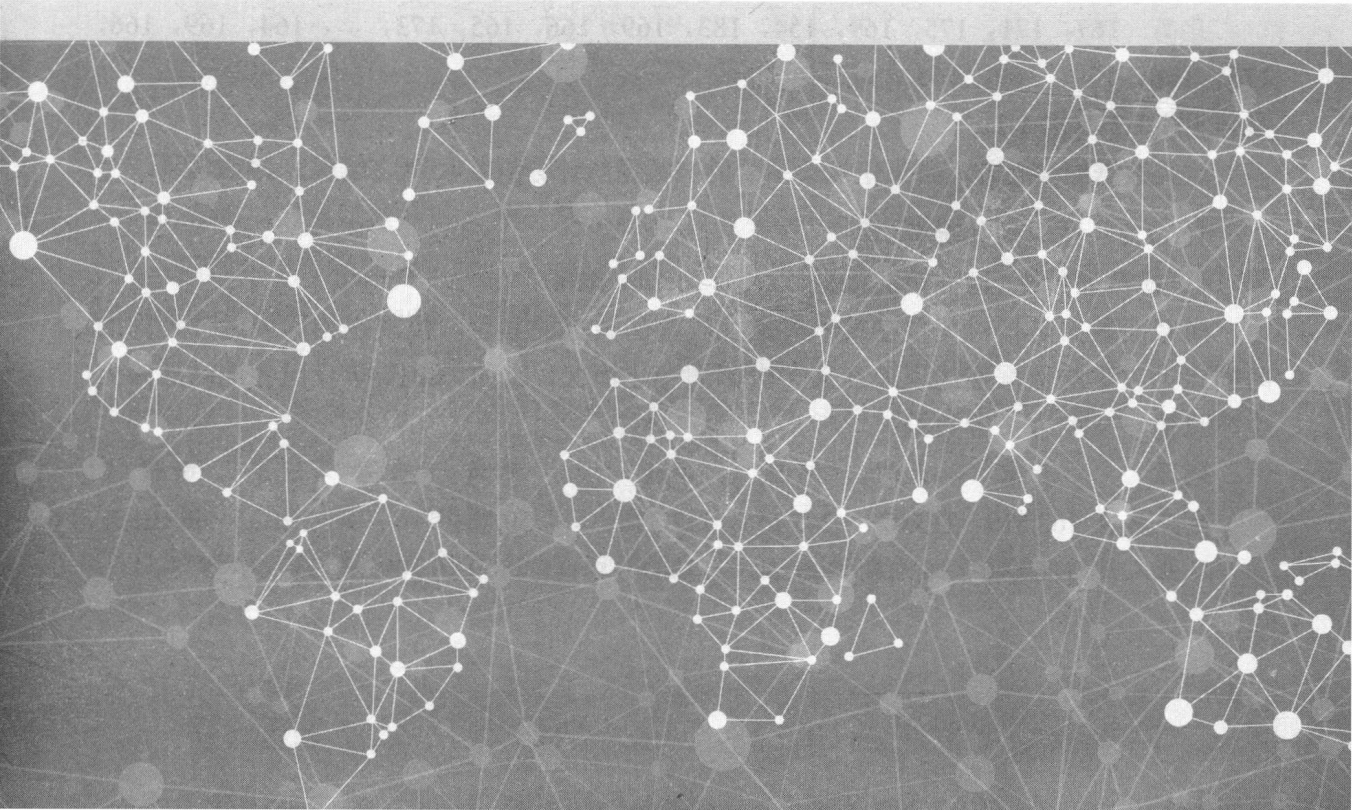
第一部分　数据分析基础知识

第二部分　数据分析高级方法

第三部分 大数据基本处理方法

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。

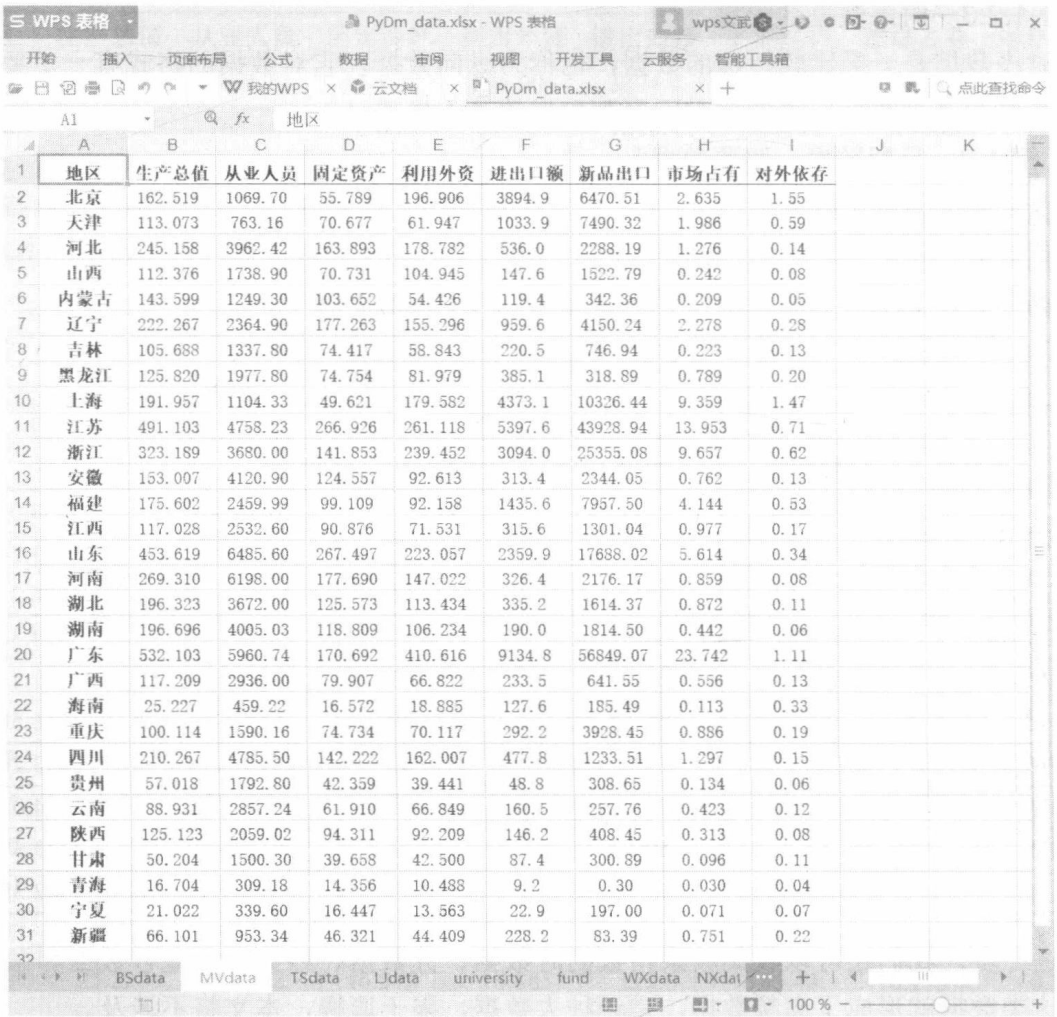
1.1.3 数据的管理

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。



	A	B	C	D	E	F	G	H	I
1	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
2	北京	162.519	1069.70	55.789	196.906	3894.9	6470.51	2.635	1.55
3	天津	113.073	763.16	70.677	61.947	1033.9	7490.32	1.986	0.59
4	河北	245.158	3962.42	163.893	178.782	536.0	2288.19	1.276	0.14
5	山西	112.376	1738.90	70.731	104.945	147.6	1522.79	0.242	0.08
6	内蒙古	143.599	1249.30	103.652	54.426	119.4	342.36	0.209	0.05
7	辽宁	222.267	2364.90	177.263	155.296	959.6	4150.24	2.278	0.28
8	吉林	105.688	1337.80	74.417	58.843	220.5	746.94	0.223	0.13
9	黑龙江	125.820	1977.80	74.754	81.979	385.1	318.89	0.789	0.20
10	上海	191.957	1104.33	49.621	179.582	4373.1	10326.44	9.359	1.47
11	江苏	491.103	4758.23	266.926	261.118	5397.6	43928.94	13.953	0.71
12	浙江	323.189	3680.00	141.853	239.452	3094.0	25355.08	9.657	0.62
13	安徽	153.007	4120.90	124.557	92.613	313.4	2344.05	0.762	0.13
14	福建	175.602	2459.99	99.109	92.158	1435.6	7957.50	4.144	0.53
15	江西	117.028	2532.60	90.876	71.531	315.6	1301.04	0.977	0.17
16	山东	453.619	6485.60	267.497	223.057	2359.9	17688.02	5.614	0.34
17	河南	269.310	6198.00	177.690	147.022	326.4	2176.17	0.859	0.08
18	湖北	196.323	3672.00	125.573	113.434	335.2	1614.37	0.872	0.11
19	湖南	196.696	4005.03	118.809	106.234	190.0	1814.50	0.442	0.06
20	广东	532.103	5960.74	170.692	410.616	9134.8	56849.07	23.742	1.11
21	广西	117.209	2936.00	79.907	66.822	233.5	641.55	0.556	0.13
22	海南	25.227	459.22	16.572	18.885	127.6	185.49	0.113	0.33
23	重庆	100.114	1590.16	74.734	70.117	292.2	3928.45	0.886	0.19
24	四川	210.267	4785.50	142.222	162.007	477.8	1233.51	1.297	0.15
25	贵州	57.018	1792.80	42.359	39.441	48.8	308.65	0.134	0.06
26	云南	88.931	2857.24	61.910	66.849	160.5	257.76	0.423	0.12
27	陕西	125.123	2059.02	94.311	92.209	146.2	408.45	0.313	0.08
28	甘肃	50.204	1500.30	39.658	42.500	87.4	300.89	0.096	0.11
29	青海	16.704	309.18	14.356	10.488	9.2	0.30	0.030	0.04
30	宁夏	21.022	339.60	16.447	13.563	22.9	197.00	0.071	0.07
31	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。