中国数据分析

CHINA DATA ANALYSIS 数据分析·因你而不凡

——中国数据分析行业核心刊物——

《中国数据分析》行业特刊
2021年第04期 总第48期(季刊)
咨询热线：400-050-6600
http://www.chinacpda.org

中国商业联合会数据分析专业委员会 主办

培养数据人才 打造"数字工匠"

大数据时代，数据成为关键生产要素，企业的数字化转型和数据应用也迈入新的阶段。数据量的爆发式增长和数据问题的暴露让越来越多的企业看到大数据的价值，将目光转向数据分析，期待从数据金矿中采炼更多，向数据驱动决策转型。然而知易行难，数据分析带给企业的也不仅仅只有机遇和收益，还有各种问题和挑战。

数据化转型是一个充分利用数据资源解决现有行业问题和社会问题的过程，这个过程不仅依赖于资金投入、资源整合，更高依赖于人才的技能、创造性以及管理能力，数据人才就是支撑企业数据化转型的重要基础和武器。如何培养数据人才、配置数据团队、制定相应的人才战略并吸引更多优秀的数据人才、激励数据人才发挥最大价值，是其成功数字化转型的关键，从而在新一轮的市场竞争中占据高地。

自 2016 年教育部第一次新增“数据科学与大数据技术”专业起，大数据相关专业新增数量在新增专业数量排行榜中均位居前列，认真分析这些专业更多体现了实用性与交叉性，反映了大数据相关专业朝着精细化、融合化的方向发展。然而，面对快速变化的商业和科技环境，学校和企业对于数据人才的培养始终面临更多形态的挑战：一方面数据人才需要大量实践经验积累和不同场景历练，才能将数据科学运用到真实的业务场景中，因此一个优秀的数据人才培养周期是非常漫长的；另一方面数据人才培养出现一定的教育偏差，数据思维能力的提升是重中之重，数据能力将是未来求职者的重要基本素质，同时企业也应鼓励人员提升数据思维能力，从提升数据质量出发从而培养企业数据文化底蕴；第三方面我国教育体制以注重培养专业化人才为主，导致现阶段既了解传统行业技术，业务流程与发展需求，又能够掌握和应用数字技术的复合型人才严重缺乏，有融合实践经验的高素质人才更是紧缺。

数据化应用在每个企业中蔓延，每个企业都无法拒绝这种趋势的影响。对于企业而言，在这个不确定、模糊、复杂、易变的商业环境中能够帮助自身应对数据化“战争”的有力武器之一就是数据人才。在数字经济发展规划中，汇聚人才要素资源，培养高精尖数据人才，打造“数字工匠”，形成一支数量充足、素质优良、结构合理、富有活力的数字人才队伍。如何最大化数据人才的价值，充分支撑企业的数据化转型，是非常值得我们探究的话题。

中国商业联合会数据分析专业委员会

本期目录 CONTENTS

主办单位
中国商业联合会数据分析专业委员会

编委成员
会员处 李苗苗

出版时间
2021 年 12 月出版 < 总第 48 期 >

美工设计
市场处 崔峻珩

联系我们
中国商业联合会数据分析专业委员会
地址 : 北京市朝阳区朝外 SOHO-C 座 9 层
电话 : 400-050-6600 / 010-5900.0991 转 652
传真 : 010-5900.0991 转 607
官网：www.chinacpda.org

欢迎广大读者踊跃投稿，内容包括学术观点、教学体验、教学活动、学习感悟、实战经验、随笔文章等。
稿件附图格式为 JPG 或 TIFF 格式，大于 1M，分辨率在 300dpi 以上。
感谢您对《中国数据分析》的支持！ 投稿邮箱：xiehui@chinacpda.org


CPDA® 数据分析师
CERTIFIED PROJECTS DATA ANALYST. SINCE 2003
从来没有一种坚持
会被辜负!
学习如同逆水行舟，
想要得到什么就要付出努力，
把一切交给时间的检验。
宜奋发　忌懈怠
www.chinacpda.com | www.cpda.cn
TEL. 400-050-6600

关于数据分析行业数据分析师事务所会员及《会员执业资质证书》的相关说明

来源 / 中国商业联合会数据分析专业委员会　编辑 / 数据委会员处　李苗苗　日期 / 2021-09

近期，中国商业联合会数据分析专业委员会（以下称“我会”）接到企业及个人反映，部分地区不断出现机构违法伪造我会颁发的《会员执业资质证书》，用以参与竞标、洽谈数据分析相关重大项目。

针对此情况，我会统一作以下说明：

1、关于我会数据分析师事务所会员及《会员执业资质证书》的说明

我会是国务院国有资产监督管理委员会、中华人民共和国民政部批准设立的专业委员会，长期指导、促进数据分析行业的行业自律。

对符合我会事务所会员要求的数据分析师事务所，依流程审批吸纳为我会会员，同时授予数据分析行业《会员执业资质证书》。每年 3 月我会对所有会员进行统一年检，年检通过即会籍有效，其相关资质可在中国数据分析行业网站（www.chinacpda.org）查询。

凡未经过我会审批颁发资质证书且无法在中国数据分析行业网站查询资质信息的单位，其所持《会员执业资质证书》均为伪造。

2、请认真甄别《会员执业资质证书》样式，利于分辨资质真伪

自 2021 年 4 月起，我会颁发的《会员执业资质证书》全面升级，新版资质证书（如图）是目前所有会员单位持有的唯一证书，新增的二维码功能具备纸质资质证书与中国数据分析行业官方网站线上电子证书一致性的查询功能，使资质证书更具防伪性，同时优化会员年检办理流程，方便广大会员进行保存和展示。

需求单位也可通过证书二维码快速查询我会会员单位从业状况、资质有效期与年检情况，有效识别伪造资质。

《会员执业资质证书》样本示意图

3、我会《会员执业资质证书》查询方式

新版证书具备线上查询资质功能，无论是会员单位还是

客户，均可通过扫描新版资质证书二维码进入或直接登录进入我会官网进行在线查询。我会会员每年须按时按要求完成年检，只有会员单位且会员单位会籍在有效期内，才能在网站查询到信息。

在线查询方式如下（如图）：

一、 登录中国数据分析行业网站：www.chinacpda.org

二、 以下两种方法可进入查询页面：

方法一：主页面悬浮框——资质查询快速通道

方法二：页面拉至左下方，从业资质查询——团体会员名称/编号 查询

三、 按图示输入纸质证中相关信息即可进行查询

4、对伪造我会颁发执业资质证书的单位，一经我会查证核实，将严厉追究其法律责任

作为中国数据分析行业专业行业组织，我会呼吁所有行业内单位严格遵守行业规定，遵循诚实信用原则，维护正常、健康的市场秩序。我会将加强行业监督力度，欢迎并鼓励社会公众监督，监督举报线索可发送邮件至：xiehui@chiancpda.org 。

随着国家大数据战略的实施，近几年各地政府和企业数字化转型需求日益增多，产生海量数据，对数据进行深层次的分析成为发展、竞争的核心。我会数据分析师事务所，在上述背景下不断发展和壮大，助力社会数字化转型加速向前。我会希望数据分析师事务所能在良性竞争的环境下发挥更多专业性，为社会提供服务，让市场看到数据分析行业更加规范，更加蓬勃！

想了解更多关于数据分析师事务所及我会会员资质信息，请登录我会官网（www.chinacpda.org）或关注我会公众号（中国商联数据分析专业委员会），同时可添加二维码咨询我会相关人员。

数据分析行业线上公益沙龙：“让你的数据会说话”

来源 / 中国商业联合会数据分析专业委员会　编辑 / 数据委会员处　李苗苗　日期 / 2021-09

9 月 9 日，中国商业联合会数据分析专业委员会特邀帆软软件九数云事业部运营总监 Jojo 先生为数据分析行业分析师带来一场主题为“让你的数据会说话——小白如何跨越数据分析学习鸿沟”的线上公益沙龙活动。

越来越多的公司开始重视数据，从经验决策转变为以数据驱动业务发展。各种岗位都开始对数据分析能力有一定的要求，比如通过数据指导运营方向、通过数据提示用户真实的产品需求；通过数据发掘销售增长点，大公司甚至会专门设置数据分析部门来专门支持业务发展。数据分析师是随着互联网时代发展起来的岗位，市场上很少有绝对对口科班出身的从业人员，多为相关专业转行而来。

现在数据分析转行热潮中，大多都是数据分析新手。他们都有一些典型常见的问题：我要学什么？我该怎么学？会一点儿 Excel 能做数据分析吗？这是因为没有体系造成的，一个人进入新领域时，如果没有全局观，很容易一叶障目，不见泰山。数据委了解数据分析岗位实际工作需求，联合九数云为分析师带来本次公益沙龙。

一个小时的活动分享，Jojo 先生从 Excel 开始的探索到瓶颈，到如何跨越数据分析的学习鸿沟，并通过实际案例项目讲解如何短期提升效率和数据思维。分析师们通过直播，直接和业内专业人士进行在线对话，启发对数据分析的深度思考与研究。

数据分析行业公益沙龙活动：5 分钟实现数据分析弯道超车

来源 / 中国商业联合会数据分析专业委员会　编辑 / 数据委会员处　李苗苗　日期 / 2021-09

9 月 23 日，中国商业联合会数据分析专业委员会特邀讲师王冲老师和赵丽老师为数据分析行业分析师带来一场主题为“5 分钟实现数据分析弯道超车”的线上公益沙龙活动。

近一个小数的分享，两位老师分别从数据分析发展应用趋势，学习路径现状到如何集中精力用思维指导业务，从 Datahoop 数据分析平台模块功能介绍到数据分析应用场景和完整案例演示。

分析师们通过直播，直接和老师进行在线对话：当下，数据整理、数据分析已经是一个企业提前规划、提前布局的参考行径，是企业切入下一赛道的征兆，是个人提高自我竞争力的机会，因此对数据分析已经到了必须了解的时候。

此次活动聚集了上百名数据分析师和数据分析爱好者聚焦直播间，点赞率过万，并且在沙龙结束后收到众多好评和希望回放的需求，深得分析师赞许。

数据分析行业公益沙龙活动：个人信息与重要数据保护

来源 / 中国商业联合会数据分析专业委员会　编辑 / 数据委会员处　李苗苗　日期 / 2021-11

11 月 10 日，中国商业联合会数据分析专业委员会特邀北京盈科律师事务所刘桂红律师为数据分析行业分析师带来一场主题为“个人信息与重要数据保护”的线上公益沙龙活动。

在“十四五”规划和 2035 年远景目标纲要草案中，对营造良好数字生态进行阐述时也明确提出要“加快快速推进数据安全、个人信息保护等领域基础性立法，强化数据资源全生命周期安全保护”。

大数据、人工智能、互联网新技术的广泛使用让隐私保护和个人信息安全问题日益突出，加强个人信息保护迫切性进一步加强。随着数据安全法的出台以及个人信息保护法的实施，无论是企业还是个人，我们应该做哪些准备。刘律师以实际案例方式，通过三个民、刑事案件的引入，进一步为数据分析师深度解读数据隐私新规下的数据使用。

“数字”东风下，行业奋发时

来源 / 中国商业联合会数据分析专业委员会　编辑 / 数据委会员处　李苗苗　日期 / 2021-10

大数据，已经是社会上最火热的行业词汇。随之而来的数据仓库、数据安全、数据分析、数据挖掘等等围绕大数据的商业价值的利用逐渐成为行业人士争相追捧的利润焦点。大数据的高速发展已应用于各个领域，包括数据分析在内的大数据垂直细分行业研究未来发展前景很大，根据 IDC 发布的有关数据预测，大数据市场规模将在 2025 年达到 19508 亿元的高点。

全球正大步迈向大数据新时代的大背景下，围绕大数据已经形成了一定的产业规模，许多企业开始参与大数据产业链，数据的存储、处理和分析等需求也越来越旺盛。

同时，我国也正在加速从数据大国向着数据强国迈进。本月初，在北京举办的“数字开启未来，服务促进发展”——2021 年中国国际服务贸易交易会大量展示了数字产业化和产业数字化方面的前沿成果和技术应用，比如，全球云计算的开创者亚马逊是如何利用云计算赋能的数字贸易、构建的智慧生活；智能汽车探索者特斯拉如何推动数字技术服务交通出行领域；中国互联网巨头阿里巴巴的数字政府、乡村振兴数字服务如何推动智慧乡村建设，等等。

国内外经济数字化发展趋势证明“数字化”已经不再是个概念，并深入到了各行业、各领域，相关产业的数据从业者肩负大任。

制造业可以利用工业大数据和数据分析提升制造业水平，包括产品故障诊断与预测、分析工艺流程、改进生产工艺，优化生产过程能耗、工业供应链分析与优化、生产计划与排程；汽车行业利用大数据和物联网技术开发无人驾驶汽车，各家“造车新势力”动态不断，无人汽车也许会在不久后走入我们的日常生活；金融业、互联网行业、餐饮行业对大数据和数据分析的应用更是数不胜数，已经与每个人深度绑定。

数据分析可以让人们对大数据产生更加优质的诠释，而具有预知意义的分析可以让数据分析师根据可视化分析和大数据分析后的结果做出一些预测性的推断。数据分析技能大概率是未来必不可少的工作技能之一。在大数据行业发展成熟的国家，有接近七成的市场决策和经营决策都是通过数据分析研究确定的。

在此背景下，我国市场也开始大量培养数据分析师，直到现在，各行业对数据分析师的需求仍然长盛不衰，而且还有扩展之势，人才缺口依旧很大。

中国商业联合会数据分析专业委员会（以下简称“数据委”）积极推动数据分析技术的普及和应用，培养专业人才，数据分析师，促进中国数据分析行业健康发展。

经国家工信部教育与考试中心和数据委授权指定的培训品牌“CPDA 数据分析师”自成立至今，结合行业发展和市场用人需求，秉承培养具有数据分析思维的数据分析师人才，使之将数据科学运用到真实业务场景中，站在业务角度用数据说话，支撑企业数字化转型，成为企业发展的关键资源，数据委授发的 CPDA 数据分析师证书已成为数据分析行业知名度高、权威性强的执业证书。

同时，数据委联手分布在全国十几个省份的数十家 CPDA 授权管理中心，上百家数据分析师事务所，培养输出了大量数据分析人才。为当地 IT、金融、医疗、零售、物流等领域的企业提供着决策支持服务，帮助行业人才进行职业规划上的全面指导，帮助广大企业推荐优秀人才，实现用人单位和行业人才间的高效匹配，优化人力资源配置。

以匠心情怀推广cpda 深耕西部人才培养沃土 优秀CPDA授权管...
2018年12月7日 作为西北地区唯一授权分中心,依托中数委的强大师资力量,以“匠心”情怀全力做好CPDA在西北地区的认证管理及推广工作,圆满完成了青海地区第一期CPDA培训考试等相...
青海新闻网 百度快照

CPDA福建省项目数据分析师受权管理中心612 - 百...
2020年1月11日 CPDA 福建省项目数据分析师授权管理中心主要机构:中国商业联合会数据分析专业委员会及工信部教育与考试中心全国授权管理中心分布 北京市授权管理中心: 地址:北...
百度文库 保障 百度快照

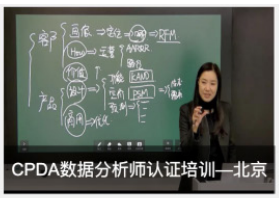
CPDA大连授权管理中心的微博_微博
CPDA大连授权管理中心 2015-10-13 14:24 来自新浪博客 发表了博文《大数据下的青岛大虾事件》从5号开始青岛大虾事件已经7天了,目测正在向纵深发展,从微博的几个关键词可以看出一些端倪。1、10月...
微博 百度快照

CPDA项目数据分析师大连授权管理中心
2021年2月14日 中设正泰数据服务(大连)有限公司——CPDA项目数据分析师大连授权管理中心,以数据服务、数据分析咨询为核心业务,以大连为中心辐射整个东北三省,重点培养数据分析...
www.antso.cn/company/cid622...asp 百度快照

大数据加持下，以往普通的工作岗位会因为数字化的加持变得更加有生命力。数据分析行业作为前述过程中的基础支撑，将会发挥至关重要的作用，会为接触、学习数据分析的企业和个人提供非常大的发展空间和发展潜力。

数据委愿和数据分析行业内外所有对行业前景看好，有实力，有意愿做出贡献和力量的企业、机构共同合作，一起努力，为实现数字化社会建设添砖加瓦。

CPDA数据分析师认证培训—山东授课现场


CPDA数据分析师认证培训—北京


CPDA数据分析师认证培训—西安


CPDA数据分析师认证培训—上海


2019年12月上海东航课程开班

数据分析行业标准化制定工作已顺利进入到征求意见稿阶段

来源 / 中国商业联合会数据分析专业委员会　编辑 / 数据委会员处　李苗苗　日期 / 2021-10

《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》的发布，让数字化转型再次迎来热议。在信息技术和数字经济的发展下，数字化转型一直是企业的关键命题，热度不减。然而，企业在数字化发展中面临的认知偏差、数据孤岛、缺乏数据管理机制和保障等问题依旧突出。

面对目前企业数字化发展问题，为了规范数据分析师事务所及大数据领域数据分析行业相关企业单位的合规经营管理，提升其竞争力和可持续发展能力；规范行业市场，监督行业自律，维护行业正当竞争，加强对数据分析从业行为的监督和指导，同时促进中国数据分析行业健康发展的需求日益增长。数据分析行业急需出台一套行之有效且务实接地气的规范性文件供从业人员、企业乃至全社会对于数据分析关注的数据分析爱好者们进行参考。

为了进一步促进国家大数据战略目标的实施，以及更好的为数据分析领域营造一个健康有序的发展平台与空间，中国商业联合会数据分析专业委员会组织开展针对《数据分析行业服务参考文件》的制定工作，自启动至今，受到社会广大大数据公司的高度重视与积极响应，得到相关政府、企业单位及专家学者的大力支持，纷纷给予积极且有效的反馈意见。

经过数月共同努力与付出，自 10 月起，《数据分析行业服务参考文件（征求意见稿）》面向社会进一步征求意见，期间得到广大媒体及行业内的宣传报道，目前征求意见稿的征求意见工作顺利进行中，收到众多企业和个人申请参与并提交建议，我会正在组织专家组对意见进行汇总和整理，希望最终形成的终稿对行业内机构的成长目标有不同阶段的路径指导支撑，同时呼吁社会各界共同关注数据分析行业发展，加强对数据分析业务的重视，共同助力企业数字化转型。

数据作为在数字经济时代的关键生产要素，其自身具有很大的经济价值，这对数据分析相关从业者来说，意味着即将迎来等待许久的发展风口和机会。《数据分析行业服务参考文件》的发布填补了数据分析行业规划范的空白，数据服务的规范化、有序化将保障数据服务的有序发展、促进以数据为关键要素的数字经济健康发展。

个人信息保护法 11 月 1 日起实施：明确不得大数据杀熟

来源 / 经济参考报　编辑 / 数据委会员处　李苗苗　日期 / 2021-11

《中华人民共和国个人信息保护法》于 2021 年 11 月 1 日起施行。法律明确不得过度收集个人信息、大数据杀熟，对人脸信息等敏感个人信息的处理作出规制，完善个人信息保护投诉、举报工作机制等，充分回应了社会关切，为破解个人信息保护中的热点难点问题提供了强有力的法律保障。

个人信息保护法共 8 章 74 条，在有关法律的基础上，进一步细化、完善个人信息保护应遵循的原则和个人信息处理规则。中国电子信息产业发展研究院网络安全所所长刘权对《经济参考报》记者表示，个人信息保护法为个人权益的保护构建了基本法律框架，也为相关个人信息处理者提供了具体的合规指引，标志着我国个人信息保护迈出了具有里程碑意义的一步，进一步完善了我国在数据领域的立法体系。

针对 App 过度收集个人信息、公共场所安装摄像头和人脸识别设备等个人信息保护中的热点难点问题，个人信息保护法给出回应，包括处理个人信息应当具有明确、合理的目的，并应当与处理目的直接相关，采取对个人权益影响最小的方式；在公共场所安装图像采集、个人身份识别设备，应设置显著的提示标识；所收集的个人图像、身份识别信息只能用于维护公共安全的目的。

针对越来越多的企业利用大数据分析、评估消费者个人特征用于商业营销的“大数据杀熟”问题，个人信息保护法给予明确禁止，规定个人信息处理者利用个人信息进行自动化决策，不得对个人在交易价格等交易条件上实行不合理的差别待遇。

专家表示，个人信息保护法将促进信息数据依法合理有效利用，为数字经济健康发展提供法律保障。中南财经政法大学数字经济研究院执行院长盘和林表示，个人信息保护法明确了个人信息的权属权益，未来互联网平台利用个人信息需要从用户获取授权，也将在使用、存储过程中承担更多信息保护的责任。

工信部对第 9992 号建议答复：持续提升工业互联网平台安全能力

来源 / 通信世界网　编辑 / 数据委会员处　李苗苗　日期 / 2021-11

为贯彻落实习近平总书记关于坚持和完善人民代表大会制度的重要思想、关于加强和改进人民政协工作的重要思想，工业和信息化部积极做好十三届全国人大四次会议代表建议、全国政协十三届四次会议提案的办理工作，特别是结合党史学习教育，切实为民办实事、解难题，强化组织指导，创新沟通机制，努力将代表委员提出的有价值、高质量建议转化为破解难题的政策措施，推动工业和信息化事业高质量发展。为深入宣传工业和信息化部 2021 年全国两会建议提案办理工作成果，工业和信息化部政务新媒体“工信微报”特开设“复文选编”栏目，陆续编发部分建议提案复文案例。

近日，工信部就田立坤代表提出的关于加强工业互联网平台安全建设的建议收悉，作出回复。

工信部称，当前，工业互联网快速发展，平台数量显著增长，融合应用日趋成熟。工业互联网平台作为工业互联网的中枢，向上承载应用生态，向下接入海量设备，面临的网络安全风险挑战与日俱增。工信部赞同田立坤代表提出的健全平台安全管理体系、提升平台安全技术防护能力、实施平台数据安全分类分级管理、加强安全检查评估等建议，将积极纳入相关工作举措。

一、已开展工作

（一）推动构建工业互联网平台安全政策体系。

一是根据《国务院关于深化“互联网 + 先进制造业”发展工业互联网的指导意见》《加强工业互联网安全工作的指导意见》等文件要求，推动构建多部门协同推进、政府监管、企业主责的安全管理格局，明确由各地通信管理局加强工业互联网平台安全监管，并对联网设备、系统进行安全监测。二是印发《关于开展工业互联网企业网络安全分类分级管理试点工作的通知》，部署启动分类分级管理试点工作，分类施策、分级防护，进一步加强平台企业网络安全管理。三是推动《工业互联网平台企业网络安全防护规范》《工业互联网数据安全防护

规范》等四项国家标准立项，加快研制工业互联网平台安全防护、安全评估、安全测试等 30 余项行业标准。

（二）强化工业互联网平台安全防护。

一是依托工业互联网创新发展工程，支持海尔、富士康等工业互联网平台企业建立安全接入、态势感知、风险预警等技术手段，建成测试验证、安全众测等多个公共服务平台，鼓励威胁诱捕、工业 APP 安全检测等安全技术产品加快突破。二是依托国家工业互联网安全技术监测服务平台，累计覆盖重点工业互联网平台百余个，持续监测和处置恶意网络行为。三是持续开展网络安全技术应用试点示范，围绕边缘层、基础设施层（云 IaaS）、平台层（工业 PaaS）、应用层（工业 SaaS）以及工业 APP 等安全防护需求，遴选平台安全防护优秀解决方案，不断加强平台安全防护能力建设。

（三）扎实推进行业数据安全管理工作。

一是坚持法律法规制度先行，积极参与《数据安全法》等法律制定工作，夯实数据安全工作法律基础。二是印发《电信和互联网行业数据安全标准体系建设指南》等文件，研究发布行业网络数据分类分级、重要数据识别等 40 余项重点行业标准。三是部署开展行业网络数据安全保护能力提升专项行动，印发《电信和互联网企业网络数据安全合规性评估要点（2020 年）》，明确合规评估指引。组织基础电信企业开展数据分类分级、重要数据识别、数据安全评估等标准贯标，指导重点互联网企业开展数据安全治理能力评估。

（四）加快建设工业互联网安全技术防护体系。

一是基本建成国家、省、企业三级协同工业互联网安全技术监测服务体系，国家平台已覆盖汽车、电子、钢铁等 14 个重要行业领域，涉及工业企业 10 万余家。二是组织开展工业互联网安全检查，开发检测工具箱和验证平台，及时发现、通报和整改安全风险隐患近 2000 个。三是依托工业互联网企业网络安全分类分级管理试点，面向平台企业开展检测评估，指导平台企业有效落实网络安全防护措施。

二、下一步工作考虑

做好平台安全保障对提升工业互联网高质量发展水平具有重要意义。下一步，我部将围绕健全平台安全管理体系、提升平台安全防护能力、强化平台数据安全保护等方面着力做好以下有关工作：

（一）健全完善工业互联网平台安全管理体系。

一是推动出台工业互联网企业网络安全分类分级管理指南，深入实施平台企业网络安全分类分级管理，强化重点平台企业网络安全管理。二是推动印发《工业互联网综合标准化体系建设指南》（2021 版），加快工业互联网平台安全分类分级管理等系列国家标准研制发布，指导企业落实网络安全主体责任。

（二）持续提升工业互联网平台安全防护水平。

一是鼓励重点平台企业建设企业级安全态势感知能力，将重点平台纳入安全监测体系，加强平台安全监测预警、应急处置。二是出台中小企业安全上云上平台政策措施，实施中小企业安全上云专项行动，为中小企业上云全流程提供安全指引。三是依托工程项目、试点示范等，进一步加大平台安全关键技术攻关和优秀项目遴选，促进网络安全技术创新应用，提升平台安全保障和服务能力。

（三）切实开展工业互联网数据安全保护。

一是落实《数据安全法》，研究制定工业互联网数据安全管理政策文件，建立健全数据分类分级保护、重要数据保护等基础制度。二是加快制定工业互联网等重点领域数据安全标准规范，组织开展标准验证和试点示范，指导企业做好数据安全保护工作。三是大力发展数据安全技术和产业，鼓励相关企业、研究机构积极参与数据可信采集、数据安全态势感知、数据溯源等数据安全关键技术研发创新和应用推广。四是组织研究数据安全保护认证体系，制定行业数据安全保护能力评估规范，建立产学研共同参与的评估认证机制，开展认证工作。

（四）系统强化工业互联网安全保障能力。

一是完善安全技术监测服务体系，扩大监测范围，丰富平台功能，提升监测质量，提高支撑政府决策、保障企业安全的能力。二是充分发挥行业威胁信息共享平台作用，推动建立跨地区、跨行业通报处置和应急联动机制，增强工业互联网重大安全风险、重大安全事件应对能力。三是健全安全检查检测机制，定期对重点平台、工业企业、工业 APP 开展检测评估，推动安全检测评估工作规范化、常态化、体系化。

迎接数字文明新时代，中国电信提速大数据应用创新

来源 / 通信信息报　编辑 / 数据委会员处　李苗苗　日期 / 2021-10

2021 全国大数据标准化工作会议近日在山东济南召开，会议发布了《数据治理工具图谱研究报告》《企业数字化转型白皮书》等研究成果，这些成果对于运营商推动大数据应用创新极为重要。中国电信认为，数字文明时代，大数据的创新和应用是核心。为此，中国电信近年不但积极推动大数据中心建设，而且持续将大数据应用于疫情防控、数字乡村治理、智慧养老、便捷就医服务等数字经济发展的方方面面。

大数据平台商业价值不断呈现

大数据时代，算力也是生产力。工信部信发司副司长王建伟在大数据标准化工作会议上表示，当前，数据已成为重要的生产要素，是加快经济发展质量变革、效率变革、动力变革的重要引擎。

大数据已形成产业规模。工信部相关负责人曾介绍，“十三五”时期，我国大数据产业年均复合增长率超过了 30%，2020 年产业规模超过了 1 万亿元人民币。

对于运营商而言，大数据的商业价值也已呈现。工信部运行监测协调局近日发布数据显示，今年 1-8 月，三家基础电信企业积极发展互联网数据中心、大数据、云计算等新兴业务，共完成新兴业务收入 1491 亿元，拉动电信业务收入增长 3.6 个百分点。其中云计算和大数据收入同比增速分别达 98% 和 34.8%。

中国电信积极打造大数据新基建

在我国，被纳入新基建的大数据中心建设不断提速。尤其是电信运营商，随着关于建设“全国一体化算力网络国家枢纽节点”的国家级战略工程——“东数西算”的正式启动，中国电信充分发挥自身云网优势，积极落实国家战略，通过数据中心积极构建“东数西算”格局，全面推进东西部地区协同发展。

6 月 30 日，总投资 102 亿的京津冀大数据智能算力中心一期项目四栋数据中心和两栋动力中心大楼完成第一批机架交付；7 月 7 日，中国电信中部云计算大数据中心（二期）建成；8 月 20 日，中国电信国家一体化大数据中心（宁夏中卫）节点正式揭牌。

据了解，我国 60% 的数据中心都由三大电信运营商持有，其中中国电信占比最高，中国联通、中国移动分别次之。

中国电信全面实施云改数转战略，一方面，加快数据中心和天翼云资源池布局，巩固内蒙、贵州两个集团级超大规模数据中心地位，形成“2+4+31+X+O”的资源布局， IDC 机架超 42 万架，遍布全国 700 多个数据中心和国内领先的基础网络能力，以及超过 300 个国内云资源池和超过 30 个海外云节点，3000 多个边缘节点，开展 MEC 建设，构建云边协同能力，

打造超强的算力基础。

另一方面，有序推动省级数据中心资源建设，积极推广成熟的创新技术应用，努力推进老旧数据中心机架资源改造升级。

除了拥有全国最大的 IDC 基础设施，中国电信还全面提供计算、存储、CDN、大数据等一揽子产品，为各级政府、各行业、大中小型企业、家庭、个人的信息化需求提供坚实承载。

中国电信重点发展大数据融合应用

大数据带动的新一代信息技术正从前沿技术变为重要应用。在福建漳州，面对厦门等地突发的疫情，中国电信快速部署大数据云平台以实现核酸检测预约；在新疆，中国电信“精准扶贫大数据平台”提升乡村治理智能化；在湖北武汉，全国医院最大规模的云数据中心——“同济云数据中心”正式启用……

中国电信董事长柯瑞文在 2021 年世界互联网大会主论坛上发言指出，人类社会正在迈向数字文明新时代。而数字文明是一种基于大数据、云计算、人工智能、物联网、区块链等新一代智能的信息通信技术，以高科技为主要特征的文明形态，核心是大数据的创新和应用，包括数据挖掘、数据互认、数据治理等，形成数字化、网络化、智能化的发展逻辑，在更高层面上促进“物质资料生产不断发展、精神生活不断丰富”。

中国电信将顺应数字文明新时代的发展趋势，坚持打造云网融合、绿色低碳的新型信息基础设施，坚持推进数字技术和经济社会发展深度融合。

2020 年中国大数据相关市场增幅领跑全球
未来 10 年数字经济渗透率将超 30%

来源 / 北京日报　编辑 / 数据委会员处　李苗苗　日期 / 2021-11

从历史上看，往往是思想革命先于科技革命，科技革命先于产业革命，产业革命先于经济全球化。思想革命上百年甚至几百年爆发一次，科技革命大部分是 60 年至 80 年爆发一次，产业革命是 40 年至 50 年爆发一次，经济全球化则是 30 年至 40 年一轮更迭。当今世界，新的思想革命、新的科技革命、新的产业革命和新一轮经济全球化正在同步发展，完全交织在一起，这在人类历史上是从来没有过的事情。

数字技术爆发式发展，成为新一轮技术革命和产业革命的主角

16 世纪至 17 世纪，世界开始了现代科学革命，科学理论和思想的革命使人类认知产生了飞跃。第一次科学革命的主导学科是力学，始于哥白尼创立太阳中心说，发展到牛顿的《自然哲学的数学原理》出版；第二次科学革命以化学原子论、生物进化论和电磁理论等的认知变革为特征；第三次科学革命以相对论和量子力学出现为标志；当前的第四次科学革命包括对物质结构的认识、对宇宙演化的认识、对生命起源的认识、对

意识本质的认识，这些重大科学问题都有原创性的突破。随后，发生了几次重大科技革命。

18 世纪以英国为代表的蒸汽机革命推动了机械化，19 世纪以美国为代表的电的发明推动了电气化，20 世纪美国发明了互联网，推动了信息化。当前新一轮科技革命袭来，世界科技发展处于快速进步之中，数字化、网络化、智能化融合发展，新技术、新业态、新产业变革加快，颠覆性技术、新兴技术不断突破，促进新经济、新动能的产生，旧的经济形态、旧的产业业态和旧的发展动能正在被新经济、新动能、新技术、新业态、新商业模式所替代。数字技术爆发式发展，成为新一轮技术革命和产业革命的主角，全球互联网、物联网、人工智能、大数据、云计算、云服务、计算机超算、机器人、3D 打印等数字技术创新，实际上是具有颠覆性的革命，不仅颠覆了人们的生产方式，还颠覆了人们的生活方式乃至生命方式。

人工智能的应用，代表了自蒸汽机发明以来最大的突破性创新，各种应用通过智能化机器学习不断优化，创新成果平等地延伸到所有经济领域、私人空间和社会生活。人工智能将由对人的体力替代转向对人的脑力替代，其推动科技、产业和社会变革的巨大潜力得到更多国家认同，美国、中国、德国、日本、韩国、俄罗斯等 16 个国家发布了国家人工智能发展战略，18 个国家正在研究制订人工智能发展计划。人工智能引发超级计算能力快速提高，据统计，截至 2021 年第二季度末，全球超大型数据中心约 659 个，相比 2016 年同期增加一倍多，未来几年会出现爆发式增长。

美国和中国继续占主要云和互联网数据中心站点的一半以上。紧随其后的是日本、德国、英国、澳大利亚、加拿大、爱尔兰和印度，合计占总数的 25%。在超大规模运营商中，亚马逊、微软和谷歌合计占数据中心足迹的一半以上，但近期表现突出的是三家中国公司——字节跳动、腾讯和阿里巴巴。数字化基础设施成为人类更高水平互联互通的新基础设施，与人工智能、超级计算能力等叠加，将出现人们此前见所未见、闻所未闻的认知革命、业态变革与生活场景。中国从 2G 跟随、3G 追赶、4G 并跑到 5G 领先，在一些领域成为科技创新的前沿部队。中国 2020 年底已建成 71.8 万个 5G 基站，目前已经超过 115 万个，2025 年将实现中国境内全覆盖。未来的世界，5G+ 强大算力，或者 6G+ 量子计算，将构成巨大的不受边界限制的无垠网络空间。

未来 30 年，数字经济将替代当前实体经济与虚拟经济两种基本经济形态

迄今为止，人类历史上最伟大的事件，就是科技革命与工业革命。马克思、恩格斯指出：“资产阶级在它的不到一百年的阶级统治中所创造的生产力，比过去一切世代创造的全部生产力还要多，还要大。”

在工业革命之前，人类发明几乎无一例外是靠长期的经验积累，有时候需要花费几代人的时间。如造纸术，公元 1 世纪中国的蔡伦发明改进造纸术，但直到 8 世纪才传到阿拉伯国家，通过阿拉伯传到了大马士革和巴格达，后来进入非洲的摩洛哥，在十一二世纪又经过西班牙和意大利进入欧洲。到 1150 年，距离蔡伦发明改进造纸术约 1000 年后，西班牙出现第一个造纸作坊；又约 100 年之后，意大利出现第一个造纸厂；再约 100 年后法国出现第一个造纸厂。可见，工业革命之前，科学技术传播路径非常狭窄，传播周期极其漫长。

近代工业革命之后，科技革命、工业革命、经济全球化加速发展，实现了从线性发展到指数级、爆发式增长。从历史维度看，欧美工业革命两个世纪的发展速度远超过前面的 2000 多年。而在中国，新中国成立后开始工业化，改革开放使中国进入现代商业文明，目前正在赶超欧美在两个世纪的发展中创造的科学革命、科技革命和产业革命成果，与发达国家并肩进入新一轮经济全球化。中国在少数领域开始“领跑”世界，在一些领域开始与西方发达国家“并跑”，在更多领域还处于“跟跑”地位。2020 年中国大数据相关市场的总体收益达 100 多亿美元，增幅领跑全球大数据市场。有报告预测认为，在 2020 年至 2024 年期间，全球大数据技术与服务相关收益将实现 9.6% 的复合年均增长率，预计 2024 年将达到 2877.7 亿美元。

以美国、日本、中国为代表的国家将科技创新立于经济社会发展的核心地位，围绕人工智能、量子信息科学、5G 等关键领域持续巩固创新。如日本制定了《科学技术创新综合战略 2020》。美国通过了《2021 美国创新与竞争法案》，加大面向未来科技创新需求的先行投资。美国政府在战略层面首次将人工智能、量子信息科学、5G 等先进通讯网络等列为国家“未来产业”。韩国发布了人工智能半导体产业发展战略。

中国信息通信研究院刚刚发布的《全球数字经济白皮书》显示，测算的全球 47 个国家数字经济增加值规模在 2020 年达到了 32.6 万亿美元，同比增长 3.0%，占 47 个国家 GDP 的比重上升到 43.7%。2020 年美国数字经济仍然蝉联全球第一，规模达到 13.6 万亿美元，全球比重高达 41.7%。中国居第二位，规模为 5.4 万亿美元。德国、日本和英国分列第三至第五。值得注意的是，德国、英国、美国数字经济在 GDP 中的占比均超过了 65%，中国比重不到 40%，但中国以 9.6% 的增速位居全球第一。

2019 年，全球服务业、工业、农业数字经济渗透率分别为 39.4%、23.5%、7.5%。未来 10 年，全球数字经济渗透率将超过 30%，未来 30 年，数字经济将替代当前两种基本经济形态——实体经济与虚拟经济，形成高渗透率的数字经济生态体系，物化的物质生产与虚拟链接、运行、存在的关联方式将再一次发生颠覆性革命。

大数据助力政治学研究

来源 / 人民日报　编辑 / 数据委会员处　李苗苗　日期 / 2021-10

随着信息技术迅猛发展，新型信息分析技术被应用到政治学研究中来。许多政治学研究者采用当前比较成熟的大数据爬取、大数据文本分析等技术，获取新的研究素材并进行分析整理。同时，以亿万为量级单位的大数据本身就具有政治学研究价值。可以说，大数据为政治学带来研究方法上的创新，也进一步拓展了政治学研究领域。

基于大数据对网络民意进行研究。网络上有许多现成数据可以用于对网络民意和政府部门回应进行研究。一些政治学研究者常把两类网络大数据用作分析素材：一类是政府网络问政平台上的群众留言及相关部门回复信息，另一类是主流媒体网站以及微博、微信等社交媒体平台上的网民发帖信息。这些数据都可通过数据爬取技术直接获取。

第一类数据经过分析处理后，可以反映群众关注的议题类型分布、政府相关部门回应的特点和问题解决程度。还可以进一步分析特定类型群体在什么时期大致提出哪些类型的诉求。

第二类数据主要用于分析网民对某一具体事件的看法和态度。对微博等社交媒体的数据分析，不仅关注具体话题，还将话题和情感、偏好等因素分析相结合，呈现网络民意对某个具体事件的态度演变过程及影响网络民意变化的具体因素。可以说，网络大数据的应用极大丰富了政治学对网络民意的研究方式。

对文本转换后的大数据进行分析。在网络大数据之后，研究者尝试挖掘其他非现成的、更具独特性的大数据信息。这类信息主要有两种，一种是从公开平台获取的大数据文本。这类数据通过数据爬取技术获得，但发布这些数据的平台并非人人皆知，需要研究者去搜寻。

目前，政治学研究者较多关注并取得研究进展的大数据主要有：从裁判文书网获取的司法大数据，从人大代表的议案建议库和政协委员的提案库中获得的文本大数据等。随着政府信息公开力度加大，这类可供研究的信息资源会越来越丰富。另一种数据需要先进行文本转换才可使用，如对历史资料的研究。相当数量的历史资料是以图片形式而非文本形式存在的，这就需要通过识别技术将图片信息转换为文字文本，然后进行大数据分析。这类数据的获取难度较高，不过一旦形成数据库，对政治学研究则有较大帮助。

大数据为政治学提供新的研究素材和分析技术，但并未改变政治学研究的本质。面对层出不穷的新素材和新工具，政治学研究者需要保持清醒头脑，不能陷入对数据和方法的盲目追求中。同时，应结合政治学研究本身的特点，发挥大数据技术优势，推动信息技术在政治学研究应用中取得更多突破。

一是开发更多样化的大数据类型。

目前，政治学研究分析的大多是文本形式的大数据。文本信息只是浩瀚信息世界中的一种，还有图片、音频、视频等大量信息类型有待开发利用。这些类型的大数据有的在商业领域已有较成熟的应用，研究者可考虑以合适方式将其应用于政治学研究。

二是探索更前沿的大数据分析技术。

现有的分析技术还离不开人工标注，需要政治学研究者从大数据中抽出很小一部分，对这部分数据进行人工阅读和标注，然后利用机器学习的方法，让计算机基于人工标注的数据去分析剩下的大量数据，尝试得出相应结论。由于人工标注的数量不等，分析的效果也参差不齐。在文本挖掘上，需要进一步开发更先进的技术。

三是实现更复杂的大数据分析目标。

现有的大数据分析主要是对数据所体现的政治现象进行描述，尚未具备解释政治现象、发现运行规律以及进行预测的功能，这需要更进一步的技术支持和研究突破。从这个角度看，大数据的开发利用在政治学研究领域还有更为广阔的发展前景。

敏感信息泄露危害大：如何从纸面到现实告别个人信息裸奔

来源 / 法治日报　编辑 / 数据委会员处 李苗苗　日期 / 2021-11

11 月 1 日起，备受关注的个人信息保护法正式施行。

网络信息时代，个人信息保护领域乱象丛生，一些企业、机构甚至个人随意收集、违法获取、过度使用、非法买卖个人信息，侵扰人民群众生活安宁、危害人民群众生命健康和财产安全。个人信息保护成为广大人民群众最关心、最直接、最现实的利益问题。

出台专门的个人信息保护法也成为近年来社会各界最为强烈的立法呼声。如何保护好个人信息、如何构筑强有力的法律威慑、如何有效遏制违法违规侵害个人信息权益的行为，是立法亟待解决的问题。经过三次审议，8 月 20 日，十三届全国人大常委会第三十次会议表决通过个人信息保护法。

作为个人信息保护领域的基础性专门法律，个人信息保护法与民法典、数据安全法、电子商务法等法律共同编织成一张个人信息保护网。值得一提的是，在全社会个人信息安全意识逐步提高的大背景下，广大人民群众对个人信息保护虽然一直保持较高的关注热情，但也普遍缺乏相关的科学知识和法律知识。个人信息保护法真正从纸面的法律条文变为手中维权的利器，并非一蹴而就。

区分敏感与非敏感信息

对于普通民众来说，哪些个人信息受个人信息保护法的保护，尤其是哪些个人信息会受到法律的特殊保护，无疑是最应该弄明白的问题。

个人信息保护法的一大亮点，是首次在法律上将个人信息区分为敏感与非敏感，采取概括加列举的定义方式，将“敏感个人信息”界定为“一旦泄露或者非法使用，容易导致自然人的人格尊严受到侵害或者人身、财产安全受到危害的个人信息”，同时对敏感个人信息的处理予以专门的、更加严格的规范。

按照个人信息保护法的规定，生物识别、宗教信仰、特定身份、医疗健康、金融账户、行踪轨迹以及不满十四周岁未成年人的个人信息等，被归类为“敏感”的个人信息。相比其他的个人信息，敏感个人信息将得到更为严格的保护。

谈及为何要对敏感个人信息提供特殊的法律保护，清华大学法学院副院长程啸指出，一方面，敏感个人信息与自然人的人格尊严、人格自由等基本权利和重大人身财产权益具有极为密切的联系。无论合法还是非法处理此类信息，都会产生重大风险甚至直接损害。

例如，掌握自然人的基因、指纹、声纹、掌纹、脸部特征等生物识别信息，就可以永久识别特定自然人。如果处理者挖空心思想着如何利用这些信息来谋取利益，那对个人可能会造成何种危险将难以预测和控制。另一方面，网络信息时代，完全禁止利用个人信息显然是不可能的，如何划定保护与合理使用的边界就成为问题核心。敏感与非敏感的区分有助于更科学地划定这一边界。此外，区分并明确列举敏感个人信息，对

于自然人、个人信息处理者以及相关职能部门来说都非常必要。

“这种区分，可以使自然人更充分意识到敏感个人信息的重要性，采取更有效的自我保护行动，更谨慎的行为，一旦发现违法行为及时举报等，也能够降低个人信息处理者履行义务的合规成本，提高对处理行为合法性的可预期性。职能部门也可以集中资源进行精准有效的执法活动，提高执法效率。”程啸说。

敏感信息泄露危害极大

敏感个人信息最核心的特点就是敏感性。

“这种敏感，就是指造成侵害或危害后果上的容易性。”程啸说，侵害或危害敏感个人信息的后果有两类，一是人格尊严受到侵害。比如，泄露个人的种族、民族、政治观点、性取向、疾病等个人信息，或者非法使用这些个人信息，会使个人遭受歧视或受到不公正的对待，这就是对人格尊严的侵害。二是自然人的人身、财产安全受到危害。比如，泄露了个人的行踪轨迹，被不法分子知悉而导致受害人被杀害；泄露银行账户信息导致银行的资金被窃取等。

尤为值得关注的是，人脸识别作为敏感个人信息的一种，一旦泄露容易对个人的人身和财产安全造成极大危害，还可能威胁公共安全，因此对其收集和使用一直广受关注。

个人信息保护法第二十六条规定，在公共场所安装图像采集、个人身份识别设备，应当为维护公共安全所必需，遵守国家有关规定，并设置显著的提示标识。所收集的个人图像、身份识别信息只能用于维护公共安全的目的，不得用于其他目的；取得个人单独同意的除外。

据悉，目前，就公共场所安装图像采集、个人身份识别设备，除了反恐怖主义法之外，尚缺乏相应的法律法规，仅有少数地方政府制定了政府规章，例如，2007 年 4 月 1 日起施行的《北京市公共安全图像信息系统管理办法》、2011 年 8 月 1 日起施行的《陕西省公共安全图像信息系统管理办法》等。但这些地方政府规章颁布的年代较早，已不适应现实要求。

鉴于此，程啸建议，个人信息保护法落地之后，应当尽快从顶层设计层面完善公共安全视频图像系统的相关法律法规，更好协调公共安全的维护与个人信息的保护。

主动拿起法律武器维权

那么，作为个人信息的权利人，该怎样做才能有效地保护好自己的个人信息尤其是敏感信息呢?

中消协近日专门给出 5 个“提醒”：

要积极学习个人信息保护法等法律规定。包括了解个人信息和敏感个人信息的处理规则、自身所享有的权利、个人信息处理者应当承担的义务以及个人信息权益受到侵害时的救济方式等。

要养成“非必要不提供”的良好习惯。除了要仔细阅读隐私协议等条款外，还要考量处理个人信息理由的充分性和提供个人信息的必要性，只在确属必要的情况下才提供个人信息或者进行授权。

要对自己授权或者提供的个人信息进行持续跟踪。不同意继续处理自己的个人信息时，要积极行使“撤回同意”权利，要求对方停止处理或及时删除其个人信息。

要注意销毁带有个人信息的单据和资料，防止因随意丢弃、使用不当等造成个人信息泄露。如妥善处理未脱敏的快递单据等带有个人信息的单据和资料，使用完后应及时销毁，或是涂抹掉关键信息后再丢弃；一些带有个人敏感信息的电子数据，如证件照片等，建议用完即删或者采用加密方式进行存储。

要主动拿起法律武器维护合法权益。当自身个人信息权益受到侵害或者发现存在违法处理个人信息行为时，要主动进行投诉、举报，提供案件线索和相关凭证，维护合法权益。

存量个人信息何去何从

伴随个人信息保护法的落地，还有一个问题值得关注，那就是存量个人信息该何去何从。

所谓存量个人信息，是指个人信息处理者在个人信息保护法实施前已经收集、存储的各类个人信息。其中，一些个人信息处理者可能会以不符合法律规定的方式收集、存储了大量的包含敏感个人信息在内的个人信息。

由于个人信息处理行为具有持续性，个人信息保护法施行后，实践中这些个人信息还可能被继续利用。“对于这种处理行为应当及时地进行法律规制。”中国人民大学法学院教授张新宝指出，由于目前还缺乏清晰的法律政策指引，完善对存量个人信息的合法合规治理是个人信息保护法落地后亟待解决的问题。

在张新宝看来，对于存量个人信息的处理，如果存在违法违规情形，其行为的性质应当如何认定需要作出司法政策上的决断。他主张，应当以个人信息保护法正式实施之日作为时间节点，区分实施前与实施后两种情形分别作出判断。

张新宝建议出台相关规章或者司法解释对这一问题作出明确规定。“如果个人信息处理者的处理活动未能达到个人信息保护法规定的规范化标准，相关职能部门应当责令其改正或者获得补充的同意，或者责令其不得进行除存储和采取必要的安全保护措施之外的处理。”

首席数据官 全球城市治理新趋势

来源 / 数据观综合　编辑 / 数据委会员处　李苗苗　日期 / 2021-10

首席数据官 (Chief Data Office，简称 CDO) 一职最早由企业创设，其主要职责是根据企业的业务需求通过数据挖掘、处理和分析，对企业未来的业务发展和运营提供战略性的建议和意见。如今，这一特殊的岗位也被应用于城市治理中，成为构建数据资源管理体系不可或缺的一环。

首席数据官制度试点“多地开花”

今年 5 月，广东省印发《广东省首席数据官制度试点工作方案》，选取省公安厅、省人社厅、省自然资源厅等 6 个省直部门以及广州、深圳、珠海、佛山、韶关、河源、中山、江门、茂名、肇庆等 10 个地市开展试点工作，推动建立首席数据官制度，深化数据要素市场化配置改革。广东试点首席数据官制度在全国属于首创，推动建立首席数据官制度，是广东省深化数据要素市场化配置改革的一项制度性安排。

随后，广州、深圳、佛山、珠海等地陆续发布首席数据官制度试点实施方案。根据《广州市推行首席数据官制度试点实施方案》，到 2022 年，将组建覆盖市区两级、市各有关部门的首席数据官工作队伍，健全首席数据官管理体系，构建权责清晰的公共数据资源开发利用制度和安全管理机制，推动公共数据资源开发利用规范化、制度化。

《佛山市首席数据官制度试点工作实施方案》要求，2021 年年底前，在各区和市重点涉及数字化管理部门试点建立首席数据官制度，明确职责范围，健全评价机制，创新数据共享开放和开发利用模式，提高数据治理和数据运营能力。2022 年 6 月底前，全面总结推行实施首席数据官制度工作，形成可复制、可推广的经验做法。

《深圳市首席数据官制度试点实施方案》提出，深圳将在市本级政府，福田等 4 个区政府，市公安局等 8 个市直单位试点设立首席数据官。深圳首席数据官的职责主要体现在六个方面：推进智慧城市和数字政府建设、完善数据标准化管理、推进数据融合创新应用、实施常态化指导监督、加强人才队伍建设、开展特色数据应用探索。

《珠海市首席数据官制度试点实施方案》提出，将用 4 个月时间在其各区、各部门中建立首席数据官制度体系。值得一提的是，该方案提出，首席数据官在本级政府或本部门信息化项目论证过程中，对项目建设是否符合数据资源治理和共享要求拥有“一票否决权”。

此外，今年 6 月，江苏省工信厅发布关于在全省推行企业首席数据官制度的通知，要求在全省建立起一支核心数字化高级人才队伍，推行企业首席数据官制度，并开展第一批企业 CDO 制度试点工作。试点企业要求，在江苏省内注册，且近三年内未出现重大违法、违规及不良信用记录的独立法人单位。同时，企业重视数据管理工作，数字化基础较好、具有较大规模数据量，或数据服务业绩突出，在本地区有一定影响力。同月，浙江省杭州市高新区 (滨江) 举行了首席数据官的授牌

仪式，确定了 58 个部门的首席数据官及联络人名单，正式开放区数字资源商店的注册申请。

政府首席数据官与企业有何不同?

过去 30 年的发展，让 CDO 的功能已经从企业内部管理制度延伸到政府政务数据管理领域。中国 (深圳) 综合开发研究院博士后工作站李恩汉在接受采访时表示，政府的首席数据官与企业是有区别的。

“目前我们看到政府关于首席数据官制度的文件，实际上应该是解决政府职能部门之间数据打通，并且加以利用的这个问题。”李恩汉说，政府始终是一个责任主体，与企业的首席数据官相比，市场考虑数据利用的利益、效率，而政府还要首先考虑责任，在开放环节最主要的应该是安全问题，但前提也是要解决“数据孤岛”的问题。与企业首席数据官相比，“政府首席数据官的角色旨在促进数据共享和透明度，提高数据驱动的决策，同时保护数据机密性和隐私。政府数据的充分利用可以增强组织绩效和成功，因此数据管理者在实现这一战略资产价值最大化方面起着至关重要的作用”。

国际上，美国是最早设立首席数据官的国家。2011 年芝加哥设立了第一位市政首席数据，2013 年在联邦政府层面任命了首位首席数据官。2019 年 1 月，特朗普总统签发的《基于循证决策的基础法案》规定“联邦政府各机构负责人应指定一名非政治任命的常任制雇员担任机构的首席数据官”。

首席数据官在政府中扮演什么样的角色?

广州大学公共管理学院教授、副院长王枫云在其发表的文章《美国地方政府首席数据官制度及其功能》中提到，首先，首席数据官必须是一种新型的复合型人才，要有强烈的大数据意识和广阔的大数据视野。其次，首席数据官要掌握最新的大数据理论、技术和方法，具有较强的大数据分析能力 (包括数据挖掘、数据存储、数据分析、数据反思和数据监控等方面的能力)。再次，首席数据官要有全面的知识结构，既要精通数据技术，又要懂得与大数据相关的政策、法规和安全等方面的知识；最后，首席数据官还要有较强的创新、组织和协调能力。

另外，在论政府首席数据官制度的建立相关研究文章中，武汉大学人文社会科学研究院副院长、武汉大学信息资源研究中心副主任夏义堃提到，设立首席数据官，统筹数据战略推进、推动政府数据资源的开放共享与开发利用已经成为许多国家政府数据治理组织体系创新的重要举措。政府首席数据官的角色旨在促进数据共享与透明度，提高数据驱动的决策，同时保护数据机密性和隐私。

在其看来，政府首席数据官的职责目标是进一步提升行政领导与业务人员对政府数据的价值认知，并将其运用到决策、流程与事务处理的优化转型上，以提高数据治理的有效性。

据介绍，越来越多的国家在中央政府层面任命各种数据主管，如首席数据官负责政府数据战略制定与数据资产管理，首席数字官负责推进政府数字化转型，首席数据分析官侧重政府数据挖掘与分析利用等。不过，专家也建议，有了首席数据官，还需要有数据治理及其战略，以推动未来政府数据分析，必要的制度设计、资源支持、条件保障以及社会合作网络等构成了政府首席数据官施展能力、发挥作用的基本生态环境。

新常态下首席数据官在企业中的五大作用

根据 Gartner 2021 年首席数据官调查显示，在开展数字化计划的企业机构中，首席数据官往往负责领导或大量参与此类计划，只有 2% 的首席数据官完全置身事外。这种情况在小型、中型、大型和跨国企业机构中十分普遍。在高级数据和分析领导者的领导或重度参与下，企业机构更有可能在创新方面表现出色并有效地创造业务价值。

随着企业适应这些不断变化的工作环境，首席数据官有了新的职责，以有效应对危机并使企业为未来的中断做好准备。以下是首席数据官在疫情之后的一些新优先事项：

▲确保业务连续性

为了让企业在疫情期间继续运作，首席数据官必须确保实时进行数据收集和分析，以便利益相关者能够做出明智的决策。此外，数据领导者必须继续重新审视业务连续性计划和核心数据平台，以确保所有数据源的可靠性和真实性。

▲确定数据保护和隐私

随着越来越多的员工继续在家工作，这种新的工作模式也使企业的关键数据资产面临新的网络攻击。此外，远程工作还存在将个人身份信息 (PII) 等敏感客户数据泄露到风险较高员工手中的风险。因此，首席数据官必须与 IT 安全团队合作以控制关键数据资产。他们还必须设置策略和权限，只允许远程员工有限地访问数据。

▲建立新的数字能力

对于首席数据官来说，新常态是构建新数字能力的最佳时机。从开发全方位客户视图和现代化数据架构到迁移到云平台，首席数据官可以采取必要的数字化举措，在应对当前危机的同时为企业的未来发展做好准备。

▲降低运营成本

精益运营正在帮助大多数企业度过疫情时代。首席数据官可以帮助企业实施这种方法。他们可以利用数据并与首席财务官合作，重新确定资源分配的优先级，并制定计划以投资新的、更安全的机会。

▲为未来的危机做准备

疫情如今敲响了警钟，要求大多数企业更加认真地对待其危机计划。作为一项新职责，首席数据官必须投资于新能力，并制定可靠的危机计划，以确保企业在未来应对类似事件时具有弹性。

隐私计算蓄力待发数据安全之路任重道远

来源 / 比特网　编辑 / 数据委会员处　李苗苗　日期 / 2021-10

在大数据技术得到普及和推广的过程中，社会大众的数据获取方式及理念发生根本转变的同时，也为个人隐私保护埋下了隐患。

据 IDC 预测，2025 年全球数据量将高 175ZB。其中，中国数据量增速最为迅猛，预计 2025 年将增至 48.6ZB，占全球数据圈的 27.8%，平均每年的增长速度比全球快 3%, 中国将成为全球最大的数据圈。

数据给企业业务发展带来商机的同时，安全、个人信息保护等话题也一直受到民众热议，“外卖、打车杀熟”、“强制刷脸”等现象已不止一次跃上热搜，而现在，上述种种现象因为国家一个法案的颁布，有望得到改善。

2021 年 8 月 20 日，《个人信息保护法》正式通过，并将于 2021 年 11 日 1 日正式施行。《个人信息保护法》的诞生，标志着我国网络数据法律体系是继《网络安全法》《数据安全法》之后，具有重要意义的一块拼图终于落定，具有划时代的意义。

比特网发现，在全球数据安全的问题越来越多地被提及，从欧盟的 GDPR，到美国的 CAPP 及其各州的法律，再到 APEC 对数据和隐私保护问题的关注，各国都将数据权利上升到前所未有的国家主权的高度。

《个人信息保护法》发布

首先最需要明确的是，《个人信息保护法》正式通过，确定个人信息保护的重要性，而且它起到了一个里程碑的作用。

其实，这不是国家首次就个人隐私问题发布政策，早在 2019 年，关于“数据隐私保护”国家层级的一系列信息安全技术规范与数据管理政策，就以前所未有的密度起草、推出和实施。

此次国家推出《个人信息保护法》，具有更强的实际意义，里面的很多条例可以最直接地保护用户。比如以往注册 APP 的时候都是首次勾选“告知通知书”便终生被授权，也就意味着我们不能撤回之前野蛮时期的任性授权便被默认为合法，这对用户权利是巨大的侵害。

《个人信息保护法》中专门对这个问题作出了新的要求，同意应由个人在充分知情的前提下自愿、明确作出，且可以被便捷地撤回。不得以个人不同意或撤回同意为由，拒绝提供产品或服务。

另外，条例还特别提到了对 14 岁以下未成年人实行更为

严苛的信息保护，这也体现出立法者对未成年人的特别关爱。

从惩罚力度上，也可以了解到国家对于整治个人隐私泄露、数据安全问题的决心。以前，很多法律条款的罚款金额都是明确的，我们发现，“情节严重的，没收违法所得，并处于5000 万以下，或者上一年度营业额 5% 以下的罚款。”

虽然，对于很多巨头公司来说，5000 万并不是一个大数字。但营业额的 5%，涉及的金额可能就会达到百亿以上，也让滥用数据的企业付出更大的代价。

大数据时代，照射不到阳光的个人隐私

如今，各种各样的移动互联网应用程序 (App) 层出不穷，极大满足了人们在社交、购物、娱乐、办公等多方面的需求。与此同时，一些手机 App 也成了个人信息泄露的“元凶”，有的“偷拍”用户人脸，有的“偷听”用户聊天，不经意间，人们的隐私信息、“网络足迹”等就被“偷走”甚至“滥用”。

目前个人数据信息泄露、过度采集、数据贩卖、数据垄断现象时有发生，最直接的后果就是“大数据杀熟”“二选一”，让你多掏钱没商量想举报数据侵权往往调查取证困难，大概率会跟平台之间妥协，至于是不是真的处理了处理好就未可知了。

当下 APP、网页、小程序等，都会要求提供用户的个人位置、身份、手机号等不同信息。各家平台都在说会对数据脱敏处理，但是你在收到快递时接到客服电话时，往往是自己真实的信息完全暴露给了别人。个人数据安全处理不好，互联网就谈不上健康发展。

而且，个人隐私泄露问题，有可能会因为数据泄露形成更大的安全问题。

2018 年 9 月，Facebook 爆出，因安全系统漏洞而遭受黑客攻击，导致 3000 万用户信息泄露。12 月，再次爆出，Facebook 因软件漏洞可能导致 6800 万用户的私人照片泄露。一系列事件影响，Facebook 股价已较当年年初下跌 29.70%。

不久前，亚马逊将因违反欧盟数据保护规则被罚款 7.46 亿欧元，约合 8.88 亿美元。这也是欧盟有史以来最大数据隐私泄露罚单。

数据泄露的事件远不止于此，此前，万豪发公告称旗下酒店喜达屋 5 亿房客信息被泄露；社交平台陌陌 3000 万用户数据在“暗网”被销售；问答网站鼻祖 Quora 遭恶意攻击，1 亿用户数据被窃；谷歌还曾因可能出现的数据泄露问题关闭旗下产品。

目前来看，企业对网络依赖越多，接入的设备越多，就越可能被黑客利用成为窃取数据的跳板，这让企业防不胜防。我们只希望，企业在把数据作为生产资料使用的同时，也应当把数据作为生产资料保护，这是需要每一家现代企业应当树立的重要观念。

虽然我国发布了《个人信息保护法》，企业自身的安全防护意识也同样重要。但是，企业从自身经营的角度数据安全这一问题到底该如何落实呢？

目前来看，隐私计算几乎是当下数据互联互通的唯一技术解决方案。隐私计算是面向隐私信息全周期保护的技术，通过对明文数据的加密，可以实现数据的可用不可见。

不过，隐私计算产品处于初步应用阶段，市场需求尚未完全挖掘。大部分行业甚至由于数字化程度低、业务流程不明确，导致缺乏市场需求。因此产品距离实现大规模工业化，仍需要进一步训练和优化。在实际运用中，技术服务平台可能只提升行业共性的业务表现，隐私计算厂商需进一步研发架构、更新底层模块，以解决个性化的业务需求。

而且，数据安全同样为隐私计算带来了前所未有的机遇，基于隐私计算“可用不可见”这一独特的优势，可以说已经成为数据互联互通的唯一技术解决方案。

我们发现，就全球范围而言，数据安全远没有达到技术的成熟期，而对于我国来说，在数据安全技术方面仍存在诸多卡脖子难题待突破，因此，在一段时间内互联网行业仍然可能存在滥用数据、泄露数据、数据贩卖等问题，当然我们也不能完全把问题推给技术，而是要从政策、经济等多方面对其进行补充。

写在最后

数据安全是新时代下的新问题，但也充满了机遇，数据互联互通是大方向，如果处理得好各生产要素之间的潜力将会被无限激发，效率也会大幅提升。但是，新型技术的应用要合理合法，无论是在人脸识别、身份认证、指纹识别等等一系列技术，跟个人信息相关的，都要适应国家的法律法规，要更加规范。

回归方程之后的一堆检验到底为啥？

来源 / CPDA 数据分析师 刘程浩　编辑 / 数据委会员处 李苗苗　日期 / 2021-10

我相信，包括以前的我在内，很多的数据分析师用统计软件拟合好回归方程后，估计扫一眼那一堆的统计检验显著性水平，只要不超过 0.1，心里那股爽歪歪的劲儿，就好比铁丝绑豆腐——甭提了！之后就屁颠屁颠开始忙着写报告了，很少去认真思考这样的一件事儿：为啥回归方程（或模型）之后要做那一堆的检验，或者这些检验的目的是什么？

不过这也不能全怪大家，因为当年我们读书的教材也没有解释为什么要这么做，只是告诉我们要做这件事儿。我们的教授也忙着搞科研项目，也没和我们讲为什么要这么做，再加上光是搞懂第一次见面的 OLS 的推导过程就已经眼花缭乱了，有多少人还有心思和精力再去补刀这个问题呢？所以我印象中相当长一段时间内，我只关注这些检验的显著性水平只要足够小，就不去管它了。

那为啥我又要去搞懂这个问题呢，其实还是工作倒逼的，说的更直接点就是客户要我用他们能听得懂的话去解释。讲真，我还真的很感谢工作以来遇到的那些客户（无论是内部客户还是外部客户），因为在他们的压力之下，我除了要把原理性的东西搞明白之外，我还得将其用对方能够听得懂的业务语言去解释，或者说业务上能够解释的清楚。这个过程虽然很辛苦，但是却不失为一种机会，只要这一关你撑过去了，那么这个模型或者方程就深深地烙印在你的脑海里了。

我们现在步入正题，来讲讲为啥回归方程之后要做那一堆的统计检验。

首先我们要搞清楚一点就是，我们为什么要做回归方程。从统计的视角来看，回归方程的作用就是将几种可观测的现象之间的相互作用，用抽象的数学方式进行显性化和固化。说的更直白些，就是要研究他们之间的规律。如果这些规律可靠或者有很高的代表性，那么我们就可以应用这个规律去做很多事情。而回归之后的这些统计检验是非常有必要的，我是举双手双脚赞成的。因为从找作用规律而言，你得排除一个很大的风险，那就是这个方程对样本的代表性不够，或者说这个方程有可能是偶然的结果。

那么回归方程之后要做哪些检验呢？我们就最主要的那 3 大检验来聊聊。

第一种统计检验，对残差的检验。

我们打个比方，采集到某个菜园大棚内一天内气温和二氧化碳浓度的数据。这个时候我们会发现温度对植物的呼吸作用会有些影响，在一定的温度范围内温度越高二氧化碳浓度会越低。进行数学建模的话，可以根据散点图的形状，做直线、抛物线、对数曲线……的回归拟合。如果我们采用的是最常用的最小二乘法进行建模的话，以线性回归为例，我们会采用这样的形式：

$$y = f(x) + \varepsilon$$

典型的回归方程，大家会联想到

$$y = ax + b + \varepsilon$$

这个时候好多人会问，怎么多出一个尾巴 ε 出来？其实，这个是有道理的。因为我们做一个直线去拟合散点图，是很难把全部点都用一条直线给串起来的，总会有些点没有在直线上。这种情况下，我们所有样本的信息就被分拆成 2 部分：一部分是可用规律来表示的，就是 y=ax+b，另一部分就是不能用这个规律来表示的，也就是误差。那么我们就专门用 ε 来表征这个误差。

但是这个误差又不能干扰整个线性方程，不然的话试想一下，如果 ε 对线性方程有影响，那么说明自变量就不止 x 一个，那么这个方程就不能叫做一元线性方程了。在上面的例子中，如果我们发现因为光照时间也会影响温度，同时光合作用也会影响二氧化碳浓度，那么，说明 x 作为自变量还不够彻底，线性方程的解释性不够充分，这样的话 ε 就包含了光照时间的影响规律在里面，就会体现出 ε“藏着其他规律”的情况。在上例中，残差和应变量 y 之间有相关关系，也会和 x 存在相关关系。

所以，为了满足上面 ε 不能与回归方程中的变量有相关关系的要求，此时 ε 就相当于的要满足以下的 3 个假设：

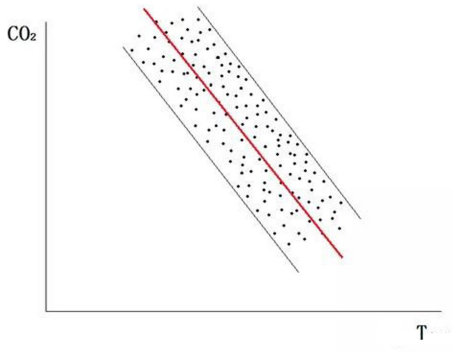
(1) 误差项 ε 是一个期望值为零的随机变量，即 E(ε) = 0。

(2) ε 的方差都相同或者固定。

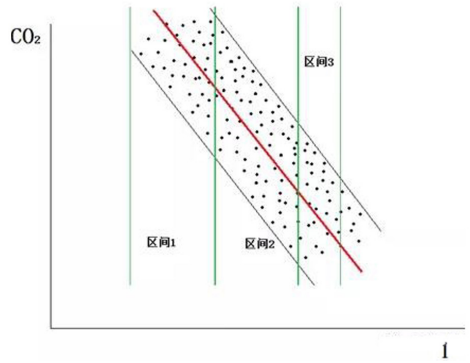
(3) 误差项 ε 是一个服从正态分布的随机变量，且相互独立。

对于第一个假设，其实很好理解，我们可以从文字和图像的角度来进行解读。首先我们看“语文视角”。我们还是以温度和二氧化碳浓度之间的关系来举例，E(ε) = 0 就意味着我们的回归方程拟合的好或足够理想，因为这说明除了温度以外的所有其他因素，它们对二氧化碳浓度所产生的总的随机扰动影响为 0。你想啊，由于总误差项（随机扰动）之和为 0，那

么必然其均值（期望）就等于 0 了。那我们再来看“美术视角”，如下图表现的那样，理想状态下回归方程拟合的好，那么这条回归直线会穿过样本点群中最具代表性的位置（满足 MSE 准则），如此一来虽然各个样本点有的分布在回归直线的两侧，有的可能就在回归直线上，但是它们各个点和回归直线之间的偏离 ε 总和却是为 0 的。如果样本采集的更凑巧一点的话，散点图和回归直线就好像一根电线的铜芯和绝缘橡胶皮的关系一样，绝缘胶皮就会扭成什么形状，铜芯就会扭成什么形状，回归方程（“铜芯”）的代表性最好。



对于第二个假设，同样也可以用“语文”和“美术”两个视角来理解。用“语文视角”来理解的话，就是无论温度越来越高 / 低，还是二氧化碳浓度越来越低 / 高，误差项 ε 都不会随之变化而变化，因为各个误差项 ε 之间都保持着固有的稳定性（方差固定）。同样用“美术视角”来看如下图，如果我们任意按照温度方向，还是二氧化碳浓度方向，将归回方程和样本点群分成若干分区间，那么每个区间里的误差项 ε 的分散程度都一样（也就是方差固定）。



对于第三个假设，则相对来说较为复杂些，但用“历史 + 语文视角”来解释可能更轻松些。关于误差项服从正态分布，得要追溯到伽利略所在的中世纪时代。那个时代科学家大都在天主教廷中任职，因为他们要观测很多天体并做研究。由于不同的望远镜，不同的观测条件，甚至不同的人用相同的正确的方法去观测天体，记录下来的结果都会存在观测误差。所以观测误差是无法避免的。但观测误差总有这样一个规律，也就是说只要不犯错误（系统性偏差），大的误差和小的误差总会出现的比较少，而大多数观测的误差都会集中在某个范围内。经过伽利略、拉普拉斯……高斯等等天文学家、数学家几百年的努力，终于被高斯证明了观测误差是一种随机扰动，服从正态分布，并给出了正态分布的数学表达式。所以，第三个假设实际上要说明的就是“只要回归方程拟合的足够理想，即把所有影响二氧化碳浓度的因素都找对了，找齐了，那么剩下回归方程和样本点之间的各个误差项 ε 就是属于一种随机扰动了”。综上所述，既然有了关于误差项的假设，我们求出了回归方程后，就得去验证下误差项是不是真的服从些这个假设。不过说实话，对误差项的检验，并不是说一定要 100% 等于这个假设，因为现实中能够完美地服从这 3 个假设的回归方程是很难找到的。但这并不妨碍我们去检验在一定的概率下，这个回归方程所产生的误差项的特征，和 3 个假设的要求有较高的一致性。

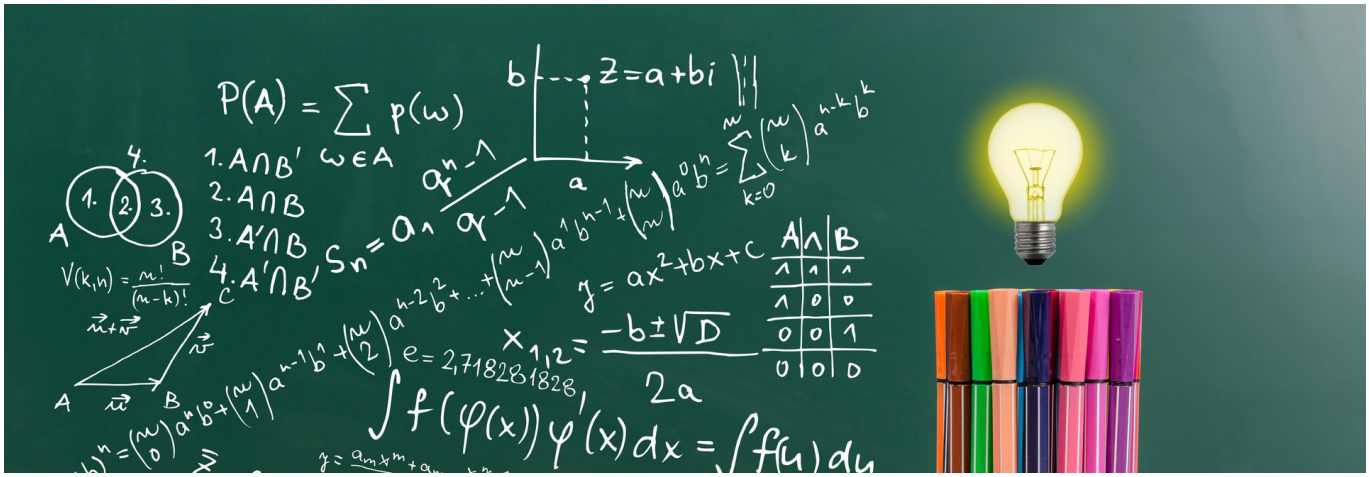
最后，ε 在实际应用中，是用残差来代替的。只要计算出了残差，我们就可以用残差去做检验。

那检验残差是不是就得对这 3 个假设一个一个去检验呢？当然你可以一个一个去检验。但其实有一个比较快速全面的检验方法。就是利用正态分布的特点，将 3 个假设合并在一起一起检验。简单的来说，就是要检验残差 e 是否服从 $N\sim(0,\ \sigma^2)$ 其中 0 代表均值，σ^2 代表固定的方差，N 就代表了正态分布。在很多统计软件里都有对残差的检验，例如 Excel 可提供残差的参数，自行进行检验；SPSS、Eviews、R、SAS……都带有这些功能。

最后再补充一点就是，如果残差检验没有通过，该怎么办呢？其实前人已经总结了很多的方法： 1. 如果只是没有通过均值为 0 的检验，但是通过了另外 2 个假设的检验。换句话说是残差分布服从，$N\sim(\mu,\sigma^2)$，$\mu\neq 0$ 那么可以对原来的回归方程增加一个截距项 μ 即可。 2. 也可以通过中心极限定理得到的“标准化残差”或根据杠杆参数换算得到的“学生化残差”，将之与自变量构成多维的散点图进行分析，推测是漏了变量还是用错了变量的表达式。 3. 有时候极端值的出现也会影响回归方程的代表性，所以也要对其进行探索式的排除。4. 还有就是社会科学领域的回归模型，可能要更多的去研究业务流程，只有对业务流程吃的比较透，才会对建模的变量和参数选择有深刻的理解，才不会那么容易遗漏变量；或者才会更快的找到被遗漏的变量。 5. 有时候残差还隐含了其他的规律，这时候对残差研究说不定还能找到其他的规律。例如著名的 ARCH（p）模型，其核心思想就是误差项在 t 时刻的方差，依赖于时刻（t-p）的误差平方大小。

$$\begin{cases} y_t = x_t'\phi + u_t,\ \ u_t\sim N(0,\ \sigma_t^2) \\ \sigma_t^2 = E(u_t^2 \mid u_{t-1}, u_{t-2}, \cdots) = \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \cdots + \alpha_p u_{t-p}^2 \end{cases}$$

第二种检验，对整个回归方程的检验：F 统计量检验。上面我们对残差做完检验后，说明我们的回归方程把样本的信息提取的是比较充分的。不过这里又冒出另外一个潜在问题，那就是这个回归方程计算结果是偶然的吗？这里可能你会很不理解：我可是花了 2 张 A4 纸，写满了 Σ 的推导公式算出来的方



程系数，为什么还会“偶然出现”呢？其实说到这里，如果不对 F 统计量做一个介绍的话，估计没法继续往下讲了。F 统计量按照解决问题的场景，是有不同的表达式的。最适合这个问题场景的表达式，应该是用于 ANOVA 分析的场景。

回忆起 ANOVA 分析，里面主要讲的是 2 组样本之间是否方差存在显著差异，从而判断 2 组是否来自同一个整体。由于研究的目的不同，我们对 F 检验的通过与否也持不同的态度。这里多啰嗦几句。如果我们是希望验证某些改良措施是有效的，那我们是希望有施加这些影响的样本，和没有施加这些影响的样本之间的结果是显著存在差异的；当然了，如果我们是希望验证某些破坏性试验构不成对原有物件的影响，那我们是希望有施加破坏性试验的样本组，和没有施加破坏性试验的样本组之间的差异是不显著的。

那么回到对回归方程的整体显著性检验来看，F 统计量该如何应用呢？回答这个问题的时候我们稍加认真思考，其实也就不难了。整个样本的信息其实被划分成了 2 个部分：被回归方程表示规律的那部分，和被残差表示没有规律的那部分。从前面残差的分析我们了解到，残差不应该和回归方程有联系，也就是说他们俩的方差应该是有显著差异的。否则，说明这个回归方程有问题：它的各个自变量系数没有使自变量的信息提取到足够让回归方程能够显著的区别于残差。因此，我们就有理由判断我们推导得到的回归方程不是必然的结果。这样，从逻辑上讲明白了为什么要做 F 统计检验的道理后，我们就可以构建 F 统计量了。首先，我是真用了 2 张 A4 纸自己独立推导出了这样的一个等式：

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

左边是观测值的总离差平方和，表示总的样本信息。等号右边第一个是代表回归离差平方和，它代表回归方程本身所产生的信息；等号右边第二个，就是我们熟悉的老朋友——残差平方和了，它代表所有非回归方程的信息。我们现在要做的就是将右边这 2 个部分的内容做一下 F 检验，看看他们是否存在显著性差异。根据 F 统计量的计算公式，还不能用他们 2 个直接比较，必须用上他们的自由度：

$$F=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2\Big/m}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\Big/(n-m-1)}\sim F(m,n-m-1)$$

在这里面，m 代表自变量的个数，n 代表样本的个数。H0 假设就是所有的自变量的“系数 i”都各自 =0，写成数学表达式就：是系数 1=0，系数 2=0……系数 n=0；H1 假设，就是“系数 i”不全为 0 那么此时，我们当然是希望 H0 假设成立啦，因此只要能够拒绝 H0 假设的话，那么就说明该回归方程的各个系数的出现不是偶然的，也有种说法是说“回归方程提取到的信息显著区别于残差的信息”；这就等价于整个回归方程不是偶然出现的了。第三种检验，对回归方程的各个系数检验，t 统计量检验。上面讲了残差的检验和对整个回归方程的检验，现在就来讲讲对回归方程的各个自变量的系数检验了。为啥还要对系数做检验呢？其实，残差的检验如果通过了，只是说明了回归方程把样本信息量提取全了，F 检验通过了只是说明回归方程不是偶然的；但这里还存在另外一个问题：是不是所有的自变量在回归方程中都是起到明显的作用，会不会有些是其实不怎么对回归方程起作用，它的引入只是起到了锦上添花的效果？比方说只是起到了从 89 分到 90 分那么点儿的效果。如果是这样的自变量，我们不考虑引入也罢。因为从回归方程的目的出发，我们是要找变量间的“重要的起作用”的变量，而不是要找这些没啥用的变量。既然如此，如何判断这些自变量是否有用呢？一个很重要的指标就是它的系数。如果某个自变量的系数太小，和 0 没有显著差异的话，那么即便给这个自变量赋予一个正常的值，那么被这个系数一乘就几乎

没作用了，因为结果也接近 0 嘛。因此，我们是不能忽视对系数的检验的。

刚才我们从逻辑上讲清楚了要对回归方程的系数要做检验，那用什么方法呢？为什么书上总是告诉我们要用 t 检验呢？为啥不用正态检验呢？

说到用 t 检验这也是有原因的。实际上我们用样本推算出来的方程系数，是一个估计值。比方说我们上面那个温度和二氧化碳浓度的例子，也就是抽取了某一天的测量值。而真正意义上大棚里的温度和二氧化碳的浓度之间的回归系数，得把大棚从正式投入生产到报废这段时间内，所有时间段、所有内部空间角落的数据都采集到（也就是总体），然后计算得到它们。

一个大棚管理的好能用个 2-3 年，然而我们不可能花个 2-3 年去采集总体数据，因此我们是希望判断这个抽样的数据算出来的回归系数，能多好地代表总体的回归系数。虽然总体的回归系数我们不可能求到，换句话说我们不知道总体的回归系数的方差是多少，但我们可以拍胸口保证总体的样本量绝对大于 30，而且远远大于 30！大到由于足够大，可以将之看成是一个正态总体，甚至某些时候直接假设总体就是服从正态分布。

当满足刚才所述：总体方差未知，但总体来自正态分布这 2 个条件时，前人通过大量实证研究得出的一个比较稳妥的方案就可以派上用场了，那就是 t 检验。t 检验之所以比较稳妥，就在于抽样的样本可能受随机因素的影响，抽取出来后并不完全呈正态性，更多的时候它呈偏态！而 t 检验能减少样本呈偏态分布产生的干扰。这就是为啥要用 t 检验的原因了。那既然选择用 t 检验，该如何操作呢？这个问题问好，我们就得从看系数的本质来回答这个问题。前面我们说到，假如推导算出的回归方程的系数很小，比方说为 0.001，还是用刚才的例子：你想即便在冬天，大棚的温度变化幅度顶多在 5-20°C之间，它再乘以 0.001 的话，就变成 0.005 到 0.02 这么小的数了。对整个方程而言影响，已经没有影响力了。

这时候你做不做检验实质上都一样。那如果我们推算出来的回归方程系数是 2.0，或者 5.5，或者 17.9……呢？这就不能像刚才那样去容易断言了。因为我们得考虑下它们出现的偶然性概率多大，如果是大概率条件下出现的，那当然要考虑接纳它；但如果它是小概率偶然出现的，我们就有理由怀疑它对变量其实是没作用的，换句话说这个系数和 0 没区别。 于是乎我们得构建若干个 H0 假设，假设这些个系数和 0 之间是不存在显著性差异的，或者直接写“系数 i=0”，然后自然就会产生若干个 H1 的假设“系数 i ≠ 0”。通过中心极限定理的推论，我们的前辈已经得出：

$$\frac{(\text{系数}_i\text{估计值} - 0)}{\text{系数}_i\text{估计值标准误差}} \sim t(n - m - 1)$$

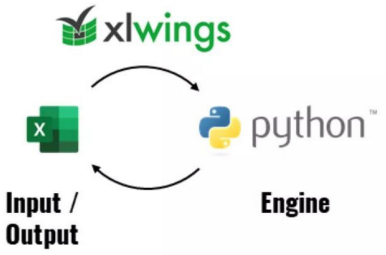
这里 n，m 的含义同 F 检验自由度的含义。这样，有了 t 统计量的代数值后，我们就可以用之来进行显著性检验了。

小结和后记这次的项目阶段性复盘的知识点，看上去应该早在校园里就该搞清楚的。然而人生总是充满了未知数。我要是早知道我工作中会用到回归建模，或者我早知道客户会来挑战我的话，我当然会在学生时代就整明白啊。不过现在搞清楚也不算晚，人生毕竟如同长跑，看的不是某个时点你跑得快还是慢，而是全程下来你到了哪个位置。或许这次复盘还漏了一些问题没搞清楚，或者说还没遇到更难对付的客户挑战，但当再遇到类似的项目时，我对建模的拿捏会更加轻松和从容。

如何在 Excel 中调用 Pandas 脚本，实现数据自动化处理？

来源 / Python 大数据分析 朱卫军　编辑 / 数据委会员处 李苗苗　日期 / 2021-10

这次我们会介绍如何使用 xlwings 将 Python 和 Excel 两大数据工具进行集成，更便捷地处理日常工作。



说起 Excel，那绝对是数据处理领域王者般的存在，尽管已经诞生三十多年了，现在全球仍有 7.5 亿忠实用户，而作为网红语言的 Python，也仅仅只有 700 万的开发人员。

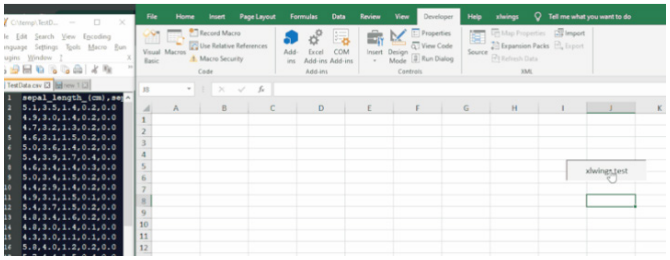
Excel 是全世界最流行的编程语言。对，你没看错，自从微软引入了 LAMBDA 定义函数后，Excel 已经可以实现编程语言的算法，因此它是具备图灵完备性的，和 JavaScript、Java、Python 一样。

虽然 Excel 对小规模数据场景来说是刚需利器，但它面对大数据时就会有些力不从心。

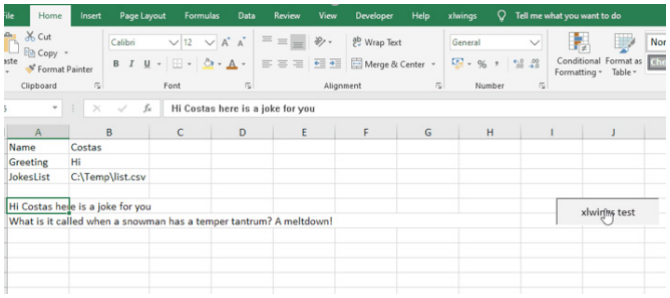
我们知道一张 Excel 表最多能显示 1048576 行和 16384 列，处理一张几十万行的表可能就会有些卡顿，当然你可以使用 VBA 进行数据处理，也可以使用 Python 来操作 Excel。

这就是本文要讲到的主题，Python 的第三方库 -xlwings，它作为 Python 和 Excel 的交互工具，让你可以轻松地通过 VBA 来调用 Python 脚本，实现复杂的数据分析。

比如说自动导入数据：



或者随机匹配文本：



一、为什么将 Python 与 Excel VBA 集成?

VBA 作为 Excel 内置的宏语言，几乎可以做任何事情，包括自动化、数据处理、分析建模等等，那为什么要用 Python 来集成 Excel VBA 呢？主要有以下三点理由：

1、 如果你对 VBA 不算精通，你可以直接使用 Python 编写分析函数用于 Excel 运算，而无需使用 VBA；

2、 Python 相比 VBA 运行速度更快，且代码编写更简洁灵活；

3、 Python 中有众多优秀的第三方库，随用随取，可以节省大量代码时间；

对于 Python 爱好者来说，pandas、numpy 等数据科学库用起来可能已经非常熟悉，如果能将它们用于 Excel 数据分析中，那将是如虎添翼。

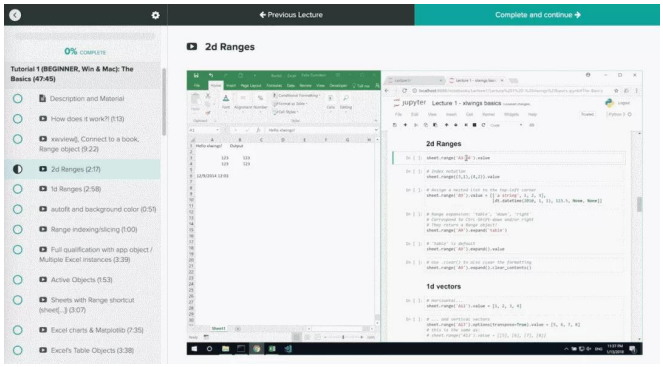
二、为什么使用 xlwings？

Python 中有很多库可以操作 Excel，像 xlsxwriter、openpyxl、pandas、xlwings 等。

但相比其他库，xlwings 性能综合来看几乎是最优秀的，而且 xlwings 可以实现通过 Excel 宏调用 Python 代码。

	可处理的Excel文件后缀		读取		写入	修改	保存	样式调整	插入图片
早起Python	.xls	.xlsx	读取	消耗时间 以10MB .xlsx文件为例					
xlrd	✔	✔	✔	12.38s	✖	✖	✖	✖	✖
xlwt	✔	✖	✖	-	✔	✔	✔	✔	✔
xlutils	✔	✖	✖	-	✔	✔	✔	✖	✖
xlwings	✔	✔	✔	7.06s	✔	✔	✔	✔	✔
XlsxWriter	✖	✔	✖	-	✔	✖	✔	✔	✔
openpyxl	✖	✔	✔	27.93s	✔	✔	✔	✔	✔
pandas	✔	✔	✔	17.55s	✔	✖	✔	✖	✖

图片来自早起 Python



xlwings 的入门使用这里不多做讲解，如果大家还不了解，先看看我之前写的入门介绍：xlwings，让 excel 飞起来！

安装 xlwings 非常简单，在命令行通过 pip 实现快速安装：

pip install python

安装好 xlwings 后，接下来需要安装 xlwings 的 Excel 集成插件，安装之前需要关闭所有 Excel 应用，不然会报错。

同样在命令行输入以下命令：

xlwings addin install

出现下面提示代表集成插件安装成功。

```
(base) C:\Users\zhuwj>xlwings addin install
xlwings version: 0.24.9
Successfully installed the xlwings add-in! Please restart Excel.
```

xlwings 和插件都安装好后，这时候打开 Excel，会发现工具栏出现一个 xlwings 的菜单框，代表 xlwings 插件安装成功，它起到一个桥梁的作用，为 VBA 调用 Python 脚本牵线搭桥。

另外，如果你的菜单栏还没有显示“开发工具”，那需要把“开发工具”添加到功能区，因为我们要用到宏。

步骤很简单：

1、在 " 文件 " 选项卡上，转到 " 自定义 > 选项 "。

2、在“自定义功能区”和“主选项卡”下，选中“开发工具”复选框。

菜单栏显示开发工具，就可以开始使用宏。

如果你还不知道什么是宏，可以暂且把它理解成实现自动化及批量处理的工具。

到这一步，前期的准备工作就完成了，接下来就是实战！

三、玩转 xlwings

要想在 excel 中调用 python 脚本，需要写 VBA 程序来实现，但对于不懂 VBA 的小伙伴来说就是个麻烦事。

但 xlwings 解决了这个问题，不需要你写 VBA 代码就能直接在 excel 中调用 python 脚本，并将结果输出到 excel 表中。

xlwings 会帮助你创建 .xlsm 和 .py 两个文件，在 .py 文件里写 python 代码，在 .xlsm 文件里点击执行，就完成了 excel 与 python 的交互。

怎么创建这两个文件呢？非常简单，直接在命令行输入以下代码即可：

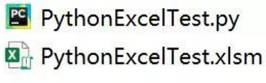
xlwings quickstart ProjectName

这里的 ProjectName 可以自定义，是创建后文件的名字。

```
(base) e:\test>xlwings quickstart PythonExcelTest
xlwings version: 0.24.9
```

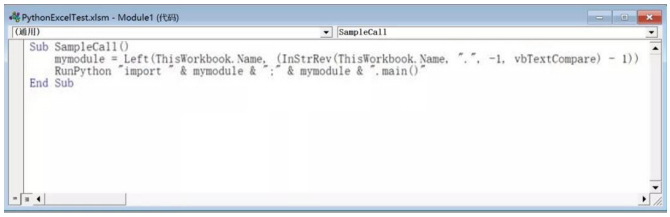
如果你想把文件创建到指定文件夹里，需要提前将命令行导航到指定目录。

创建好后，在指定文件夹里会出现两个文件，就是之前说的 .xlsm 和 .py 文件。



我们打开 .xlsm 文件，这是一个 excel 宏文件，xlwings 已经提前帮你写好了调用 Python 的 VBA 代码。

按快捷键 Alt + F11，就能调出 VBA 编辑器。




```
Sub SampleCall()
    mymodule = Left(ThisWorkbook.Name, (InStrRev(ThisWorkbook.Name, ".", -1, vbTextCompare) - 1))
    RunPython "import " & mymodule & ";" & mymodule & ".main()"
End Sub
```

里面这串代码主要执行两个步骤：

1、在 .xlsm 文件相同位置查找相同名称的 .py 文件

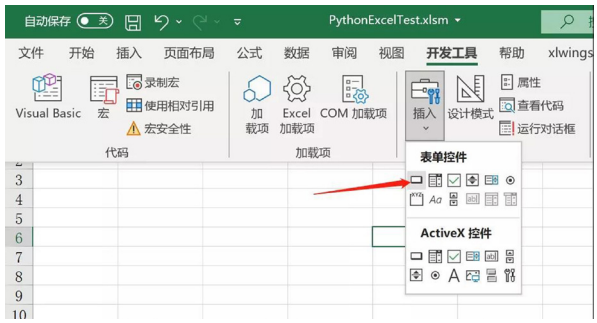
2、调用 .py 脚本里的 main() 函数

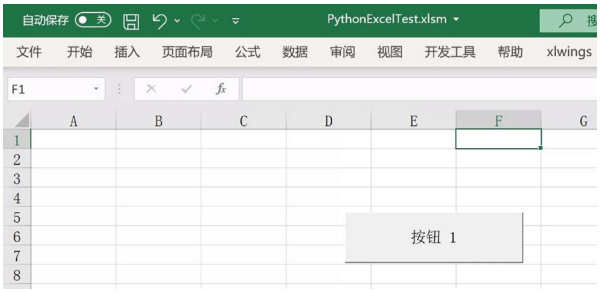
我们先来看一个简单的例子，自动在 excel 表里输入 ['a','b','c','d','e']

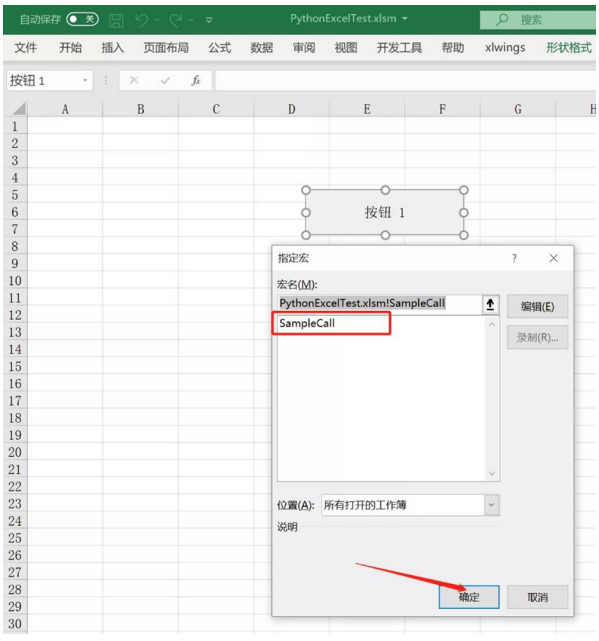
第一步：我们把 .py 文件里的代码改成以下形式。

```
import xlwings as xw
import pandas as pd
def main():
    wb = xw.Book.caller()
    values = ['a','b','c','d','e']
    wb.sheets[0].range('A1').value = values
@xw.func
def hello(name):
    return f"Hello {name}!"
if __name__ == "__main__":
    xw.Book("PythonExcelTest.xlsm").set_mock_caller()
    main()
```

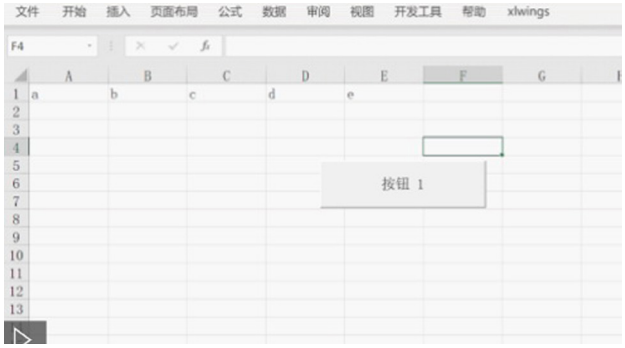
然后在 .xlsm 文件 sheet1 中创建一个按钮，并设置默认的宏，变成一个触发按钮。





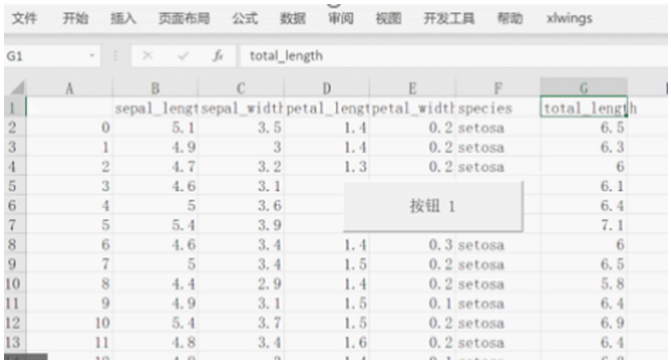


设置好触发按钮后，我们直接点击它，就会发现第一行出现了 ['a','b','c','d','e']。



同样的，我们可以把鸢尾花数据集自动导入到 excel 中，只需要在 .py 文件里改动代码即可，代码如下：

```
import xlwings as xw
import pandas as pd
def main():
    wb = xw.Book.caller()
    df = pd.read_csv(r"E:\\test\\PythonExcelTest\\iris.csv")
    df['total_length'] = df['sepal_length'] + df['petal_length']
    wb.sheets[0].range('A1').value = df
@xw.func
def hello(name):
    return f"Hello {name}!"
if __name__ == "__main__":
    xw.Book("PythonExcelTest.xlsm").set_mock_caller()
    main()
```



好了，这就是在 excel 中调用 Python 脚本的全过程，你可以试试其他有趣的玩法，比如实现机器学习算法、文本清洗、数据匹配、自动化报告等等。

Excel+Python，简直法力无边。

参考 medium 文章

如何做好数据分析？你需要这个思维框架

来源 / 一个数据玩家的自我修养　作者 / GClober　编辑 / 数据委会员处　李苗苗　日期 / 2021-11

编者荐语：

做好数据分析，除了掌握工具、理论和业务，还需要具备什么素质呢？

以下文章来源于一个数据玩家的自我修养，作者 GClover

人人挂在嘴边的数据分析，到底包含哪些方面？学好 Python 真的就能做好数据分析吗？

数据分析，拆开来看其实是几个方面：工具、理论、业务、工具，指的是我们从事数据分析所使用的具体工具，如 SQL、Excel、Python、R、SAS 等；

理论，指的是我们从事数据分析时所依赖的理论基础，如概率论、统计学、机器学习及相关的建模和分析框架；

业务，指的是数据分析落地的具体场景，输入和输出以及要解决的具体问题。

工具和理论都是比较容易速成的，这也是为什么各类网课主要集中在这些领域。业务是依赖于在行业的经验，因此，转行最好先在同行业里面转，可以借用之前对于行业的业务理解，快速上手。

以上三个方面固然重要，但并不是数据分析的全部。

数据玩家还想再加一个维度，就是思维模式。

也就是，我们除了数据分析的工具、理论以及业务知识，还需要具备数据分析的思维。那么什么叫做数据分析思维呢？我认为可以分为三个方面：

第一、定量思维

迪斯尼通过草坪规划道路的故事大家也许都听过：

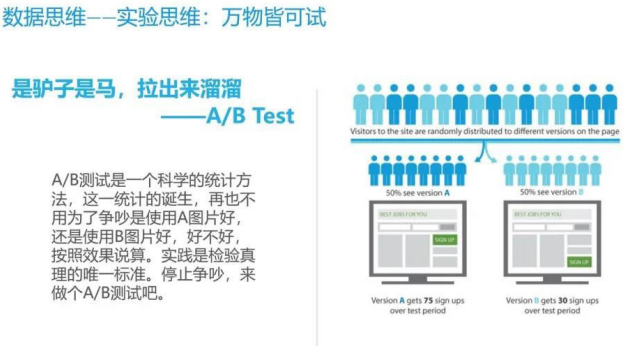
在迪斯尼乐园提前开放的半年里，草地被踩出许多小道，这些踩出的小道有宽有窄，优雅自然。第二年，格罗培斯让人按这些踩出的痕迹铺设了人行道。1971 年在伦敦国际园林建筑艺术研讨会上，迪斯尼乐园的路径设计被评为世界最佳设计。后来，迪斯尼还推出了 MagicBand，这个手环可以在园内支付，可作为酒店房卡，可以用来当 FastPass，可以用来停车等等，通过这些环节收集的数据，就可以知道哪几个项目最热门，哪几个项目不太热门，什么位置餐厅人满为患，说明还需要增加配置，什么地方餐厅无人问津，可能要做优化……等等，时间一长，积累的数据就有了各种价值，看起来无法测量的东西，通过巧妙的收集数据，都可以测量。这就是数据思维第一条，万物皆可测。

第二、相关思维

大数据时代，随着算力的不断加强，原来小样本的计算已经可以升级为全样本计算，并且可以发现变量间的相关关系，用来代替原来小样本中推导出的因果关系。

最经典的例子就是 08 年的 Google Flu：Google 流感趋势（Google Flu Trends，GFT）是 Google 于 2008 年推出的一款预测流感的产品。Google 认为，某些搜索字词有助于了解流感疫情。Google 流感趋势会根据汇总的 Google 搜索数据，近乎实时地对全球当前的流感疫情进行估测一个搜索行为，和一个疾病的发生，看似不相关的两件事情，存在强相关，这在原来是不可想象的。不过，尽信数据不如无数据，一定要找到业务含义。就拿 Google Flu 来说，在研究成果公布以后，研究人员发现结果不再准确了。经过反复确认和调研，发现因为很多人得知了这项成果，抱着好奇的心态尝试搜索关键字——尽管他们周围并未出现相关病例，导致预测结果不再准确。当你观测的对象知道你在观测他的时候，观测结果就不再准确了。

第三、实验思维



告别拍脑袋决策，告别依赖个人审美决策，告别依赖个人经验决策，通过实际的数据表现来决策。同时，根据实验结果不断的迭代和优化模型。当然，实验的前提是测量，必须先将所有实验的数据采集下来，才能根据实验数据进行决策，同时，根据数据分析的结果，可能某些人群针对某个方案更加有效，这又会用到相关思维，即某些要素的相关性决定了最后的数据表现。通过以上三个思维模式，我们可以将实际中的业务问题进行拆解，转化为数据分析问题。这么说可能还是比较抽象，具体来看看如何应用。

在广告营销领域，有一个著名的说法：

这是相当长的一段时间，广告营销行业最大的痛点，蒙着眼睛放广告，来了客户也不知道是广告带来的，还是自己找上门来的，或者其他渠道推荐来的。那么，用上数据分析思维的广告营销，会变成什么样子呢？

运用定量思维，那就是营销效果要可以度量。一个广告投出去，我需要知道到底带来了多少转化，每个渠道的转化率怎样，以及这些客户的后续活跃程度如何，是不是假量？是不

是羊毛党？是不是僵尸户？等等。

那么如何度量呢？我们自然可以想到，要检测转化率，那就要对每个渠道进来的客户打标签，定期出报表，监控每个标签下客户的活跃情况等等，自然的就形成了客户分群经营，分群营销，分群活动投放等等策略。

运用相关思维，那就是通过相关性分析，使得广告的投放更加精准。减少无效的广告投放，在更相关的人群上投放他们感兴趣的广告，提升转化率，节省营销费用。那么如何进行相关性分析呢？

通过前期采集的数据，使用 Apriori 、Collaborative Filtering 等算法，找出用户特征、用户行为及其最终购买之前的相关关系，从而优化投放及推荐模型。运用实验思维，那就是通过实验，判断哪个投放模型更优，哪个投放渠道更优，同时根据反馈不断迭代和优化模型。那么如何进行实验呢？自然是通过 A/B Test 方法，随机均分流量到不同的投放模型上，同时采集客户的反馈，不断的根据反馈迭代和优化模型。

总的来说，做好数据分析，除了掌握工具、理论和业务，还需要具备数据分析的思维，有了数据分析的思维框架，更容易将业务、理论和工具贯通，形成自己的数据分析框架，更好、更有效的进行数据分析工作。

数据体系和专题分析实战

转自 / 数据不吹牛　来源 / 产品遇上运营　作者 / 小 z　编辑 / 数据委会员处　李苗苗　日期 / 2021-11

一、互联网公司数据职能设置

互联网公司普遍十分重视数据，数据部门职能设置却各不相同。大多会设置独立的 BI 部门（如携程、京东），有些（如亚马逊）也会把数据人员分散在各个团队。

数据职能常见的有三个主要角色：

a. 数据工程师，负责搭建底层数据架构，定义数据埋点规范、编写埋点代码（有时也会由开发人员植入埋点代码）、以及建立和管理数据库报表。

b. BI，负责根据业务需求在数据库中抓取对应数据项，编写 SQL 代码，生成各类报表。（注：传统的数据库管理员（DBA）的职能更类似于数据工程师 + BI – 埋点）

c. BA，负责对 BI 生成的报表进行分析，结合业务知识对数据进行透彻解读，输出有明确指导意义的观察和建议。

BA 人员通常需要有较强的业务背景知识，能够准确地理解数据背后的业务状况和波动原因，并用业务“语言”输出分析结论。我在实践中的体会是：两种组织架构方式各有明显的利弊，优缺点截然相反。当数据人员集中在一个部门时，数据库管理和报表定制均十分专业高效。但因为离业务部门较远，业务理解受到影响，在数据定义和解读上相对偏薄弱。

数据职能分散在各个业务线时，正好相反。并有较严重的数据重复拉取，人力浪费不说，还因口径定义上的差异，导致同一数据在不同部门各不相同。例如转化率 = 订单数／访客数，有的部门在访客数中去除“疑似机器人”部分，有的部门则统一访客数为“二跳访客”，带来转化率数据的明显差异。一个比较好的做法是把数据工程师和 BI 集中在数据部门，在各个业务线分别设置 BA 人员，两边对接。

二、数据使用方式

互联网需要进行数据观察的领域十分广泛，每个细分领域都有不同的核心 KPI，应当根据核心目标拆分背后的影响因素，有针对性地提出数据需求，制定数据报表。通常数据的使用方式分为如下情况：

1. 常规数据报表

常规数据报表主要用于需要长期持续观察的核心数据。例如：

· 流量漏斗监控，可分为首页跳失率、商详页到达率（分为浏览 - 商详、搜索 - 商详两大分支）、加车率、结算率、结算完成率等核心环节漏斗数据。

· 用户渠道来源情况，如各渠道来源的用户数、新客数、订单占比、转化情况等等。

· 品类转化率波动，如各品类的流量、订单、SKU 销售数量等。

· 流量分发效率，如各频道／栏目的 CTR、商详页到达、

转化、复访率等。

当常规监控的核心数据项发生超阈值波动或趋势性波动时，通常会触发专题分析，并根据分析结果采取相应对策，以推动数据回到常规范围。常规数据报表建议通过公司的 BI 系统定制在线报表，按监控频度进行观察分析。

2. 专题分析

专题数据分析通常按专题的主要影响因素确定数据项，拆分观察维度，抓取多维度数据，对某个专题目标进行分析，找到影响因素所在的数据维度，得出结论，指导后续动作。例如：

· 针对某个重大事件的状况或效果分析，如双 11 大促后的数据总结盘点。

· 核心数据出现重大波动，如 Web 平台转化率持续提升的原因分析。

· 出现趋势性状况，如某付费渠道来源的用户数量持续下降。

· 某个专题研究，如 95 后导购特征和消费特征分析。

3. AB 测试

产品经理常有的困惑是，当上线了某一个功能或者频道后，目标数据出现了某种变化。然而，变化背后的影响因素非常多，例如时间因素导致的差异（如工作日的转化率高于周末）、竞争对手的动作、季节性因素等等。核心数据的波动往往是这些影响因素综合作用的结果，很难准确界定该功能本身带来了多少直接影响。运营也常有类似的诉求，例如当首页图标做了飘红，或者引导文案做了一些调整，数据出现了波动，但却很难确定多大程度为该特定运营动作的效果。

上述情况下，最好的方法就是做 AB 测试：取两个数据集，在数据集样本的选取中对各种影响因素做均匀的随机分布（如地域、用户群体特性），并对其中一个数据集实施特定产品功能或运营动作；在同一时段中，观测目标数据在两个测试集上的差异，从而精确判定待观测功能／动作的准确效果。这里要特别注意两点：1. 为了确保统计效果的准确性，需要有较大的样本量和统计时长（结果数量 = 用户量 * 统计时长，要么用户量足够大，统计周期可以略短；如果用户量较小，则需要更长的统计周期）。2. 如果某一个样本中存在少数对均值影响巨大的样本（例如一个金额巨大的订单），需要予以排除，以减少偶然性带来的偏差。

4. 个性化

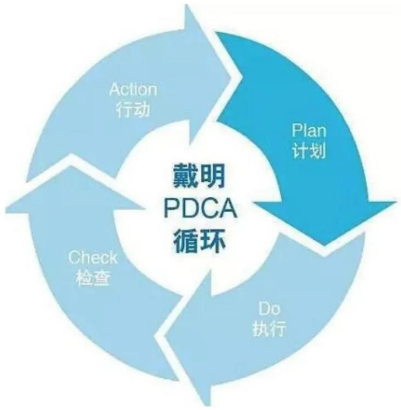
这是个大数据的时代，差异巨大的用户群体面对海量的商品和选择，“千人一面”带来的糟糕体验已不再适用。每个用户在系统中都会留下自己的线索和足迹，体现自己在商品品类、价格段、品牌偏好等方面的阶段性需求。系统可以通过数据有效发现当前用户的当前需求，进行有效的推荐，而用户也会感受到系统“懂我”，产生良好的购物体验。亚马逊早年的“Everything Store”理念，在当前时代下，也逐渐转化为“Everyone Store”，也就是我们常说的“千人千面”。数据是千人千面的基础，通过机器学习和算法设计，让系统在各个模块中进行智能化推荐，自动组装匹配当前用户的场景，是数据使用的最重要方式之一。这部分我会在后续文章中结合实际案例重点展开。

三、常规性数据报表的定制及数据监控

为了最优使用 BI 资源并突出自身专注点，在定制常规性数据报表时，切勿大而全。需要完全考虑清楚的主要有两点：北极星指标、指标监控频度。

1. 北极星指标

任何一个业务要能不断优化和提升，做出更好的效果，都需要正确设立核心指标，持续监控，并根据实际数据与阶段性预期进展之间的差距进行分析，触发相应的调整动作，以使得业务的发展和计划保持一致。这套思路在项目管理理论中被总结为 PDCA，即计划（Plan）、执行（Do）、校验（Check）、响应（Act），在项目管理和持续质量改善中也被称为戴明循环。该体系是业务目标管理的核心方法，感兴趣的同学可以查阅项目管理理论，本文不进行赘述。



从 PDCA 概念中可以看到，目标的制定、执行成效的判断以及纠偏动作的效果，都需要好的数据指标进行衡量，并作为最终目标达成与否的判断依据。这个可度量的指标，与目标呈直接的正相关关系，该指标被称为北极星指标。北极星指标体系通常分为多级，每一级指标的设立选取，都是为了更好的支持上一级指标的达成，以最终共同实现公司顶层战略（公司级的北极星指标）。在这里举个实际例子。一个电商公司的经营规模往往通过公司的年营业额（GMV）来衡量，也即 GMV 是整个公司的北极星指标之一。营业额有多种拆分计算方式，在此列出常见的一种简化计算方式：GMV = AC * Freq * Conversion * AOS

· AC：活跃顾客数

· Freq：顾客平均访问频度

· Conversion：转化率

· AOS：平均单均价

上面四个核心指标，则为第二级核心指标，通常可下达到各个部门分别负责。例如，市场部负责流量和用户数及其活跃度，产品和运营负责转化率指标，类目线负责单均价指标。

于是这些指标成为各个部门的北极星指标。如果一个指标的核心影响因素分散在多个部门，也由同一个部门牵头负责。为了达到上述各个二级指标，还可以进一步拆分。以活跃顾客数为例：AC = RC + NC – EC

· RC：留存顾客数

· NC：新客数

· EC：流失顾客数

于是这些指标又可以进一步分配到负责拉新和留存的职能团队，成为这些团队的北极星指标，由这些团队各自牵头负责。负责拉新的团队，又可以进一步把拉新指标拆分到渠道，如付费渠道、免费渠道等，进行下一级的核心指标定义和目标制定。同样地，下一级负责付费渠道的职能团队或人员，则可以进一步拆分到具体渠道，如网盟、SEM、应用商店等，进一步制定各个渠道的具体目标。如此层层往下，直到直接可控的最下一层。以此类推，产品和运营负责的转化率指标，则可以沿转化漏斗拆分为首页到商详、搜索到商详、商详加车率、购物车结算率、支付成功率等，通过逐层递进的拆分具体到各个团队进行分解，成为各自的北极星指标。

对于各个职能部门／团队来说，自己所负责的这一级指标以及下一级指标情况，应当成为常规数据报表的监控内容，由此制定报表格式，向 BI 部门提出数据需求。站在宏观维度来看，三级指标的达成可以确保二级指标的达成，二级指标的达成可以确保顶层指标的达成，从而为业务目标提供保障。因此，指标体系的合理拆分和严密监控纠偏对公司目标实现至关重要。

2. 指标监控频度

常规数据报表的周期通常为日报、周报、月报、季报。实时数据监控通常为应急响应需要（如故障宕机、突发事件处理），而半年报、年报则大多为业务结果的统计，周期过长，发现的问题及响应过慢，通常不在常规数据报表的范围。每个业务单元都具有各不相同的特点，需要进行有针对性的数据统计频度设定。下面以产品和运营层面对转化率的监控为例：

· 实时监控

在大促期间观察活动效果，流量变化迅速，高峰此起彼伏，爆品库存时有告罄，此时数据观察应当精确到最小颗粒度甚至实时监控数据曲线，对数据体现的问题（如售罄、宕机、技术故障、黄金资源位单品滞销、页面陈列错误、价格设置错误导致的波动等）迅速响应，优化促销品及资源位，并使用赛马机制，调整会场流量分发，以把大促效果推到极致。

· 日报表

对于日常促销活动，可以以天为单位，对促销品类和促销方式在整体转化漏斗中的表现进行观察，定位问题点并迅速进行针对性优化；如换品，换促销规则，更新活动页／活动栏目，配置促销标签等，以达到最佳活动效果。

· 周报表

运营方面，例如首页或频道运营，可以以周或月为单位，通过各板块 CTR、停留时间、商详到达率、加车率、转化率、复访频度等维度观察栏目用户的兴趣指数，对于薄弱环节通过数据进行深入分析（如用户动线跟踪、区域点击热度分析、跳失分析等），并适当结合用研的定性定量深访对频道入口交互设计、页面信息架构设计、频道子栏目铺设、信息展示、营销文案等进行优化，以达到最佳效果。

· 月／季报表

移动时代受到移动端发包频度的限制（大多为每两周到一个月发一个包），高度依赖技术功能的核心指标往往以月或季为单位进行统计。例如，对于核心转化漏斗模块的功能迭代和新产品模块的效率效果，可以以月或季为单位（与技术发版周期和新栏目用户教育养成周期有关），结合季节性因素，纵向对比同比和环比相应数据的波动，找到可以发力优化提升的环节。

运营动作一般带来较快速的数据响应，侧重于日报、周报对运营的指导；而产品动作一般受技术发版影响，数据响应周期适中，更偏重月或季为周期的报表，但都谋求发现问题后迅速响应。年报总体来说可能更适用于公司战略和业务线的财务考量，除了成果和得失总结，产品和运营侧的使用相对较少。

上述是针对转化率的举例。如果是用户运营和增长，同样可以根据频度对用户的渠道来源和激活情况、传播效果（短周期，如天或周）、活跃度、品类渗透率、交易情况、人均价值（中周期，如月）、留存率、流失返回率、生命周期情况（长周期，如季或半年／年）进相应的数据报表制定和监控，并触发响应的调整动作。最后，在报表制定时，建议不要把太多级别的数据放在同一个报表上，造成数据的汪洋大海，表格过度复杂，也会迷失专注点。通常一个报表含两级指标为最佳。例如，一级指标的报表只含一、二级指标数据，对于一级指标的波动从二级指标进行观察，找到波动原因。如果需要继续深入，建议另外定制二级指标报表，含二、三级指标数据。以此类推。

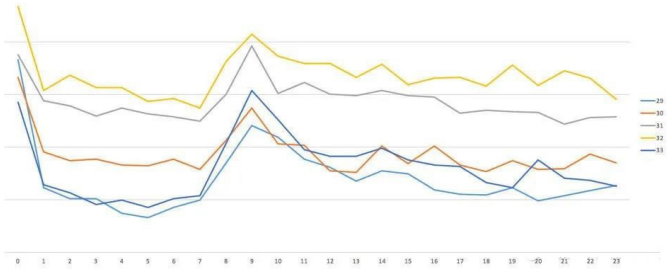
四、专题分析

工作中常会碰到一些突发异常情况，例如某阶段用户转化率大幅波动、交易金额飙升或锐减、某栏目 CTR 暴跌等，再或者观察到某些趋势性的变化（如消费者导购偏好演变、品牌消费趋势变化）。此时通常会进行专题性分析，以明确下一步解决问题的思路。

1. 专题分析触发原因

专题分析主要由如下情况触发：

a. 在数据报表中，我们常常看到一些核心数据指标产生波动，当波动范围超过一个预定义的警戒阈值时，就应该触发分析（无论正向的还是负向的波动），以理解波动背后的原因，并采取相应的对策。多大幅度的波动值得触发分析因指标本身特性对应的业务敏感度而定。阈值设置没有固定规则，大家可以根据影响的承受力来设定。这里有一个常见错误，就是对正常的小幅波动太过敏感，触发频繁的分析，最终却没有有价值的发现，属于自然波动，浪费了人力。什么是正常幅度的波动，可以对一个大时间段的同一指标进行同比环比的统计后判断。



例如，上图是我们在某五周期间观察到到流量按时间段到分布情况。大家仔细看下有什么异常？猜对了，0 点出现大流量！ 9 点，14 点，19 点的流量峰值符合移动端用户在早晨通勤时间、下午回到座位、傍晚通勤时间的访问规律。但 0 点出现如此之大的流量，十分异常，就应当触发专题分析。

b. 在数据报表中，数据体现出某个同趋势性的连续变化，例如，连续 7 次正向或负向的增长。此时，即使还没有达到预设的异常警戒阈值，都应当进行分析，以理解趋势背后的原因。可能有同学会问，为什么是 7 次呢？

其实这不是绝对的，当一个连续趋势出现时，同向的数据点越多，表明背后有某种非偶然因素的可能性越大。从统计学角度，如果是偶然因素导致连续 7 个点往同一个方向发展，可能性只有 1/128，大约为 8%。因此，7 点同趋势变化背后存在非偶然因素的置信度已经足够高了。如果是特别关键的指标，连续 5 个点同向发展（97% 的确定性）也许就该进行分析了。想要深入了解的同学可以搜索“7 点原则”，查阅 PMP 或者统计学有关的理论知识。

当然，背后应当去除已经理解的影响因素，例如越来越靠近春节时流量持续下滑，或者接近换季时新一季的服装销售持续上升，都是正常现象，除非波动过大严重脱离同比情况，否则这样的趋势并不值得浪费人力进行分析。

c. 对某个数据背后的原因感兴趣，需要分析和理解该数据背后蕴含的信息。这个和数据的波动本身没有关系，只是深入去理解数据背后的原因或因素。例如，分析为什么在平台上第三方商家的流量达到 48%，以制定更平衡的流量分发策略来扶持自营或第三方业务；分析为什么付费渠道来源的用户占比偏低或单客成本过高，以做更精准更高性价比的流量采买投放。

2. 专题分析常用方法

简单概括，专题性分析的主要做法是，按多个维度全面对波动数据指标的下层构成进行拆分，观察对比各个下层数据，找到在哪个细分维度出现异常波动，并锁定该维度，层层递进，深入分解，直到最终找到答案。在拆分到下层维度过程中，需要考虑从多个角度出发，反复对比。

例如，如果某一周发现转化率产生异常波动，可以按如下维度进行拆分观察：

维度一：商品品类拆分到各个品类，观察是否由某个品类的转化率大幅波动带动了整体转化率的波动。

案例 1：某一周我们发现全站转化率飙升近 2%，通过二级报表对各品类转化率进行观察后发现，转化率波动主要出现在美妆品类。进一步对美妆品类各 SKU 的销售进行观察，发现洁面仪、水牙线、和某款面膜等三个商品短时间销量巨大。这三个单品的上线价格远比京东和天猫更为低价，并与市场部确认，市场部有在“什么值得买”网站进行投放，导致大量用户涌入，销量激增，通过这三个热销爆款的销售推动了全站转化率的波动。

案例 2：有一次服装线的采销对某品牌服装在设置促销券时忘记设置互斥，导致用户可以反复领券和叠加用券。而该技术漏洞被人在乌云平台所披露，导致大规模的用户和黄牛涌入抢购，零元购买，极短的时间里卖出数千件，造成转化率瞬时飙升。因为人工设置价格和促销时错误难以绝对避免，此类问题在各个电商平台时有发生。

维度二：用户群体拆分到各个用户群体，观察是否由于某个用户群体的购买情况变化造成了转化率的波动。注意用户本身就可以按很多个维度拆分：

· 性别
· 地域：省、地区
· 消费价格段：高、中、低价格段
· 消费风格类型：例如时尚人群，母婴人群，数码控，阅读爱好者，家庭主妇……

案例 3: 某一周的数据观察中我们发现全站转化率的飙升，通过地域和品类的分析，发现是由于华东地区高温，导致空调风扇等商品在华东的销售飙升，推高全站转化率。北京地区雾霾爆表也曾导致净化器、口罩等商品在北京地区销售猛增。

维度三：渠道来源拆分到各个用户来源渠道，按渠道对应的销售情况进行观察。例如，有时转化率大幅提升，分析发现是因为市场部在某些导购网站的黄金资源位进行了爆款投放，从该渠道产生了巨大的流量和销售进而推高了整体转化率。当然部分渠道的刷单现象也常常会引起整体转化率波动。

维度四：转化漏斗观察首页到商详，商详到购物车，购物车到结算，结算到支付等转化漏斗环节的细分转化率的变化情况。

案例 4：有一周转化率低于警戒值，通过漏斗分析发现支付环节成功率大幅下滑。对支付渠道进行分解后发现某银行渠道的支付成功率下降到零。与该银行沟通后确认，该银行对支付接口进行了升级，升级版本存在问题，导致该支付渠道支付失败，导致整体转化率产生波动。

案例 5：有一次技术团队上线新版本后，发现转化率下跌，通过漏斗分析发现，在新用户注册环节有较大的注册成功率下降。进一步通过注册流程的分析，看到产品功能上增加了一步强制实名认证，导致部分用户在这一步由于各种考虑而放弃了注册。在与产品经理沟通后把实名认证改为可跳过，改为在后续阶段进行引导认证。这一步改变使注册成功率得以恢复，问题解决。

维度五：设备平台观察 iOS，Android，PC，Web 等各个平台以及各个 app 版本的转化率情况。例如，我们有时发现，新发的 Android 包存在技术故障，导致用户大规模登录失败，进而影响整体转化率。

维度六：销售渠道很多平台会对接下一级分销渠道，各个渠道的销售情况变化也会带来整体转化率波动。有时某个渠道进行了效果极佳广告投放，会重大促进该渠道的销售，进而影响整体转化率。

维度七：流量或销售时段分布拆分到各个用户来源渠道，按渠道对应的销售情况进行观察。例如，有时转化率大幅提升，分析发现是因为市场部在“什么值得买”的黄金资源位进行了爆款投放，从该渠道产生了巨大的流量和销售进而推高了整体转化率。当然部分渠道的刷单现象也常常会引起整体转化率波动。

案例 6：有一次转化率下降报警，数据分析表明销售情况在用户、渠道、品类等方面都分布均匀。最后产品经理与 BA 联合排查，发现在 0 点到 7 点之间有大流量出现，并且流量集中在整点刚到时爆发，由此基本可以推测这些流量并非真实顾客，而是某种程序脚本整点触发导致。最后与技术团队跟进分析，确认是某搜索引擎爬虫开始集中爬取平台商品、价格信息。

维度八：用户账号或商户有时某个商户，或某些用户，出现异常大规模订单，导致整体转化率、单均价等出现巨大波动（此类现象往往是刷单导致）。通过按商户或用户账号的销售情况拆分，可以发现此类问题。

在我和数据团队所做过的实际的分析中，以上八种维度都经常发现问题。并不排除还有更多维度，大家可以按自己的业务特性进行类推。以上只是对转化率进行分解分析的一个例子。任何一种指标通常都可以向下拆解，直到最后发现问题所在，而上面列举的八个维度，通用于绝大部分的线上状况分析。

具体的做法是：按各个维度对指标拆分到下一级后，观察下级各维度指标是否均匀体现该波动。如果是，则基本可以排除是该维度的因素所导致。对同级的各个维度逐一拆分观察，通常会发现某个维度下的某个次级指标剧烈波动，锁定该指标，再次对其下层指标进行分解观察，层层递进，最终可以找到结论。

商品零售购物篮分析实战案例：Apriori 关联规则算法

来源 / 犀数院　编辑 / 数据委会员处　李苗苗　日期 / 2021-10

购物篮分析是通过发现顾客在一次购买行为中放入购物篮中不同商品之间的关联，研究顾客的购买行为，从而辅助零售企业制定营销策略的一种数据分析方法。

本案例使用 Apriori 关联规则算法实现购物篮分析，发现超市不同商品之间的关联关系，并根据商品之间的关联规则制定销售策略。

目标

通过对商场销售数据进行分析，得到顾客的购买行为特征，并根据发现的规律而采取有效的行动，制定商品摆放、商品定价、新商品采购计划，对增加销量并获取最大利润有重要意义。请根据提供的数据实现以下目标：

1. 构建零售商品的 Apriori 关联规则模型，分析商品之间的关联性。

2. 根据模型结果给出销售策略。

分析方法

购物篮关联规则挖掘的主要步骤如下：

1. 对原始数据进行数据探索性分析，分析商品的热销情况与商品结构。

2. 对原始数据进行数据预处理，转换数据形式，使之符合 Apriori 关联规则算法要求。

3. 在步骤 2 得到的建模数据基础上，采用 Apriori 关联规则算法调整模型输入参数，完成商品关联性分析。

4. 结合实际业务，对模型结果进行分析，根据分析结果给出销售建议，最后输出关联规则结果。

数据探索分析

查看数据特征以及对商品热销情况和商品结构进行分析

1. 数据特征

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

In [2]:

```
inputfile = '/home/kesci/input/data_act_combat5529/GoodsOrder.csv' # 输入的数据文件
data = pd.read_csv(inputfile,encoding = 'gbk') # 读取数据
data .info() # 查看数据属性
print("-"*40)
print(' 描述性统计结果：\n',data.describe().T) # 输出结果
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43367 entries, 0 to 43366
Data columns (total 2 columns):
id 43367 non-null int64
Goods 43367 non-null object
dtypes: int64(1), object(1)
memory usage: 677.7+ KB
----------------------------------------
描述性统计结果：
count mean std min 25% 50% 75% max
id 43367.0 4908.589504 2843.118248 1.0 2455.5 4828.0 7380.5 9835.0
```

In [3]:

```
data.head()
```

Out[3]:

(如右图所示)

	id	Goods
0	1	柑橘类水果
1	1	人造黄油
2	1	即食汤
3	1	半成品面包
4	2	咖啡

2. 分析热销商品

销量排行前 10 商品的销量及其占比

In [4]:

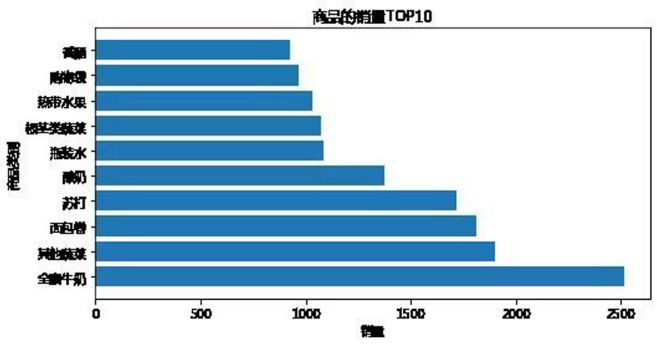
```
group = data.groupby(['Goods']).count().reset_index() # 对商品进行分类汇总
group_sorted = group.sort_values('id',ascending=False)
print(' 销量排行前10商品的销量:\n', group_sorted[:10]) # 排序并查看前 10 位热销商品
```

```
销量排行前 10 商品的销量 :
Goods id
7 全脂牛奶 2513
8 其他蔬菜 1903
155 面包卷 1809
134 苏打 1715
150 酸奶 1372
99 瓶装水 1087
```

70 根茎类蔬菜 1072
85 热带水果 1032
143 购物袋 969
160 香肠 924
画条形图展示出销量排行前 10 商品的销量

In [5]:

```
x=group_sorted[:10]['Goods']
y=group_sorted[:10]['id']
plt.figure(figsize = (8, 4))
plt.barh(x,y)
plt.xlabel(' 销量 ')
plt.ylabel(' 商品类别 ')
plt.title(' 商品的销量 TOP10')
plt.savefig('./top10.png')
```



In [6]:

```
# 销量排行前 10 商品的销量占比
data_nums = data.shape[0]
for idnex, row in group_sorted[:10].iterrows():
print(row['Goods'],row['id'],row['id']/data_nums)
```

全脂牛奶 2513 0.05794728710770863
其他蔬菜 1903 0.0438812922268084
面包卷 1809 0.04171374547466968
苏打 1715 0.039546198722530956
酸奶 1372 0.031636958978024765
瓶装水 1087 0.025065141697604168
根茎类蔬菜 1072 0.024719256577582033
热带水果 1032 0.023796896257523
购物袋 969 0.022344178753430026
香肠 924 0.021306523393363617

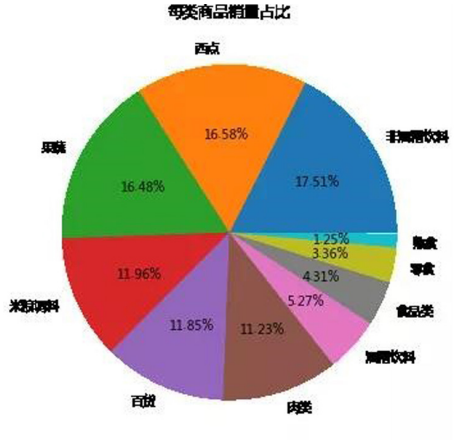
In [7]:

```
inputfile1 = '/home/kesci/input/data_act_combat5529/GoodsOrder.csv'
inputfile2 = '/home/kesci/input/data_act_combat5529/GoodsTypes.csv'
# 读入数据
data = pd.read_csv(inputfile1,encoding = 'gbk')
types = pd.read_csv(inputfile2,encoding = 'gbk')
group = data.groupby(['Goods']).count().reset_index()
sort = group.sort_values('id',ascending = False).reset_index()
data_nums = data.shape[0] # 总量
del sort['index']
# 合并两个 datafreame,on='Goods'
sort_links = pd.merge(sort,types)
# 根据类别求和，每个商品类别的总量，并排序
sort_link = sort_links.groupby(['Types']).sum().reset_index()
sort_link = sort_link.sort_values('id',ascending =
```

False).reset_index()
del sort_link['index'] # 删除“index”列
求百分比，然后更换列名，最后输出到文件
sort_link['count'] = sort_link.apply(lambda line: line['id']/data_nums,axis=1)
sort_link.rename(columns = {'count':'percent'},inplace = True)
print(' 各类别商品的销量及其占比 :\n',sort_link)
# 保存结果
outfile1 = './percent.csv'
sort_link.to_csv(outfile1,index = False,header = True,encoding='gbk')
各类别商品的销量及其占比 :
Types id percent
0 非酒精饮料 7594 0.175110
1 西点 7192 0.165840
2 果蔬 7146 0.164780
3 米粮调料 5185 0.119561
4 百货 5141 0.118546
5 肉类 4870 0.112297
6 酒精饮料 2287 0.052736
7 食品类 1870 0.043120
8 零食 1459 0.033643
9 熟食 541 0.012475

画饼图展示每类商品销量占比

In [8]:
data = sort_link['percent']
labels = sort_link['Types']
plt.figure(figsize=(8, 6))
plt.pie(data,labels=labels,autopct='%1.2f%%')
plt.title(' 每类商品销量占比 ')
plt.savefig('./persent.png') # 把图片以 .png 格式保存



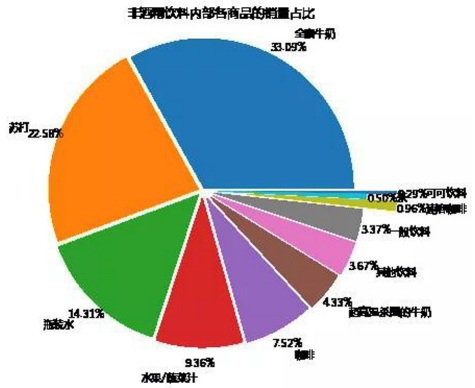
通过分析各类别商品的销量及其占比情况可知，非酒精饮料、西点、果蔬 3 类商品的销量差距不大，占总销量的 50% 左右。

进一步查看销量第一的非酒精饮料类商品的内部商品结构，并绘制饼图显示其销量占比情况

In [10]:
# 先筛选“非酒精饮料”类型的商品，然后求百分比，然后输出结果到文件。
selected = sort_links.loc[sort_links['Types'] == ' 非酒精饮料 ']
# 对所有的“非酒精饮料”求和
child_nums = selected['id'].sum()
# 求百分比
selected.loc[:,'child_percent'] = selected.apply(lambda line: line['id']/child_nums,axis = 1)
selected.rename(columns = {'id':'count'},inplace = True)
print(' 非酒精饮料内部商品的销量及其占比 :\n',selected)
outfile2 = './child_percent.csv'
sort_link.to_csv(outfile2,index = False,header = True,encoding='gbk') # 输出结果
非酒精饮料内部商品的销量及其占比 :
Goods count Types child_percent
0 全脂牛奶 2513 非酒精饮料 0.330919
3 苏打 1715 非酒精饮料 0.225836
5 瓶装水 1087 非酒精饮料 0.143139
16 水果 / 蔬菜汁 711 非酒精饮料 0.093627
22 咖啡 571 非酒精饮料 0.075191
38 超高温杀菌的牛奶 329 非酒精饮料 0.043324
45 其他饮料 279 非酒精饮料 0.036740
51 一般饮料 256 非酒精饮料 0.033711
101 速溶咖啡 73 非酒精饮料 0.009613
125 茶 38 非酒精饮料 0.005004
144 可可饮料 22 非酒精饮料 0.002897

In [11]:
# 画饼图展示非酒精饮品内部各商品的销量占比
data = selected['child_percent']
labels = selected['Goods']
plt.figure(figsize = (8,6))
# 设置每一块分割出的间隙大小
explode = (0.02,0.03,0.04,0.05,0.06,0.07,0.08,0.08,0.3,0.1,0.3)
plt.pie(data,explode = explode,labels = labels,autopct

= '%1.2f%%',
pctdistance = 1.1,labeldistance = 1.2)
# 设置标题
plt.title(" 非酒精饮料内部各商品的销量占比 ")
# 把单位长度都变的一样
plt.axis('equal')
# 保存图形
plt.savefig('./child_persent.png')



通过分析非酒精饮料内部商品的销量及其占比情况可知，全脂牛奶的销量在非酒精饮料的总销量中占比超过 33%，前 3 种非酒精饮料的销量在非酒精饮料的总销量中的占比接近 70%，这就说明大部分顾客到店购买的饮料为这 3 种，而商场就需要时常注意货物的库存，定期补货。

数据预处理

前面对数据探索分析发现数据完整，并不存在缺失值。建模之前需要转变数据的格式，才能使用 Apriori 函数进行关联分析。这里对数据进行转换。

In [12]:
inputfile='/home/kesci/input/data_act_combat5529//GoodsOrder.csv'
data = pd.read_csv(inputfile,encoding = 'gbk')
# 根据 id 对“Goods”列合并，并使用“，”将各商品隔开
data['Goods'] = data['Goods'].apply(lambda x:','+x)
data = data.groupby('id').sum().reset_index()
# 对合并的商品列转换数据格式
data['Goods'] = data['Goods'].apply(lambda x :[x[1:]])
data_list = list(data['Goods'])
# 分割商品名为每个元素
data_translation = []
for i in data_list:
p = i[0].split(',')
data_translation.append(p)
print(' 数 据 转 换 结 果 的 前 5 个 元 素：\n', data_translation[0:5])

数据转换结果的前 5 个元素：

[[' 柑橘类水果 ', ' 人造黄油 ', ' 即食汤 ', ' 半成品面包 '], [' 咖啡 ', ' 热带水果 ', ' 酸奶 '], [' 全脂牛奶 '], [' 奶油乳酪 ', ' 肉泥 ', ' 仁果类水果 ', ' 酸奶 '], [' 炼乳 ', ' 长面包 ', ' 其他蔬菜 ', ' 全脂牛奶 ']]

模型构建

本案例的目标是探索商品之间的关联关系，因此采用关联规则算法，以挖掘它们之间的关联关系。关联规则算法主要用于寻找数据中项集之间的关联关系，它揭示了数据项间的未知关系。基于样本的统计规律，进行关联规则分析。根据所分析的关联关系，可通过一个属性的信息来推断另一个属性的信息。当置信度达到某一阈值时，就可以认为规则成立。

Apriori 算法是常用的关联规则算法之一，也是最为经典的分析频繁项集的算法，它是第一次实现在大数据集上可行的关联规则提取的算法。除此之外，还有 FP-Tree 算法，Eclat 算法和灰色关联算法等。本案例主要使用 Apriori 算法进行分析。

模型具体实现步骤 :

· 设置建模参数最小支持度、最小置信度，输入建模样本数据

· 采用 Apriori 关联规则算法对建模的样本数据进行分析，以模型参数设置的最小支持度、最小置信度以及分析目标作为条件，如果所有的规则都不满足条件，则需要重新调整模型参数，否则输出关联规则结果。

目前，如何设置最小支持度与最小置信度并没有统一的标准。大部分都是根据业务经验设置初始值，然后经过多次调整，获取与业务相符的关联规则结果。本案例经过多次调整并结合实际业务分析，选取模型的输入参数为：最小支持度 0.02、最小置信度 0.35。其关联规则代码如代码所示。

In [13]:
from numpy import *
def loadDataSet():
return [['a', 'c', 'e'], ['b', 'd'], ['b', 'c'], ['a', 'b', 'c', 'd'], ['a', 'b'], ['b', 'c'], ['a', 'b'],
['a', 'b', 'c', 'e'], ['a', 'b', 'c'], ['a', 'c', 'e']]

```
def createC1(dataSet):
C1 = []
for transaction in dataSet:
for item in transaction:
if not [item] in C1:
C1.append([item])
C1.sort()
# 映射为 frozenset 唯一性的，可使用其构造字典
return list(map(frozenset, C1))
# 从候选 K 项集到频繁 K 项集（支持度计算）
def scanD(D, Ck, minSupport):
ssCnt = {}
for tid in D: # 遍历数据集
for can in Ck: # 遍历候选项
if can.issubset(tid): # 判断候选项中是否含数据集的各项
if not can in ssCnt:
ssCnt[can] = 1 # 不含设为 1
else:
ssCnt[can] += 1 # 有则计数加 1
numItems = float(len(D)) # 数据集大小
retList = [] # L1 初始化
supportData = {} # 记录候选项中各个数据的支持度
for key in ssCnt:
support = ssCnt[key] / numItems # 计算支持度
if support >= minSupport:
retList.insert(0, key) # 满足条件加入 L1 中
supportData[key] = support
return retList, supportData
def calSupport(D, Ck, min_support):
dict_sup = {}
for i in D:
for j in Ck:
if j.issubset(i):
if not j in dict_sup:
dict_sup[j] = 1
else:
dict_sup[j] += 1
sumCount = float(len(D))
supportData = {}
relist = []
for i in dict_sup:
temp_sup = dict_sup[i] / sumCount
if temp_sup >= min_support:
relist.append(i)
    # 此处可设置返回全部的支持度数据（或者频繁项集的支持度数据）
    supportData[i] = temp_sup
    return relist, supportData
    # 改进剪枝算法
    def aprioriGen(Lk, k):
    retList = []
    lenLk = len(Lk)
    for i in range(lenLk):
    for j in range(i + 1, lenLk): # 两两组合遍历
    L1 = list(Lk[i])[:k - 2]
    L2 = list(Lk[j])[:k - 2]
    L1.sort()
    L2.sort()
    if L1 == L2: # 前 k-1 项相等，则可相乘，这样可防止重复项出现
    # 进行剪枝（a1 为 k 项集中的一个元素，b 为它的所有 k-1 项子集）
    a = Lk[i] | Lk[j] # a 为 frozenset() 集合
    a1 = list(a)
    b = []
    # 遍历取出每一个元素，转换为 set，依次从 a1 中剔除该元素，并加入到 b 中
    for q in range(len(a1)):
    t = [a1[q]]
    tt = frozenset(set(a1) - set(t))
    b.append(tt)
    t = 0
    for w in b:
    # 当 b（即所有 k-1 项子集）都是 Lk（频繁的）的子集，则保留，否则删除。
    if w in Lk:
    t += 1
    if t == len(b):
    retList.append(b[0] | b[1])
    return retList
    def apriori(dataSet, minSupport=0.2):
    # 前 3 条语句是对计算查找单个元素中的频繁项集
    C1 = createC1(dataSet)
    D = list(map(set, dataSet)) # 使用 list() 转换为列表
    L1, supportData = calSupport(D, C1, minSupport)
    L = [L1] # 加列表框，使得 1 项集为一个单独元素
    k = 2
    while (len(L[k - 2]) > 0): # 是否还有候选集
```

```
    Ck = aprioriGen(L[k - 2], k)
    Lk, supK = scanD(D, Ck, minSupport) # scan DB to get Lk
    supportData.update(supK) # 把 supk 的键值对添加到 supportData 里
    L.append(Lk) # L 最后一个值为空集
    k += 1
    del L[-1] # 删除最后一个空集
    return L, supportData # L 为频繁项集，为一个列表，1，2，3 项集分别为一个元素
    # 生成集合的所有子集
    def getSubset(fromList, toList):
    for i in range(len(fromList)):
    t = [fromList[i]]
    tt = frozenset(set(fromList) - set(t))
    if not tt in toList:
    toList.append(tt)
    tt = list(tt)
    if len(tt) > 1:
    getSubset(tt, toList)
    def calcConf(freqSet, H, supportData, ruleList, minConf=0.7):
    for conseq in H: # 遍历 H 中的所有项集并计算它们的可信度值
    conf = supportData[freqSet] / supportData[freqSet - conseq] # 可信度计算，结合支持度数据
    # 提升度 lift 计算 lift = p(a & b) / p(a)*p(b)
    lift = supportData[freqSet] / (supportData[conseq] * supportData[freqSet - conseq])
    if conf >= minConf and lift > 1:
    print(freqSet - conseq, '-->', conseq, ' 支 持 度 ', round(supportData[freqSet], 6), ' 置信度：', round(conf, 6),
    'lift 值为：', round(lift, 6))
    ruleList.append((freqSet - conseq, conseq, conf))
    # 生成规则
    def gen_rule(L, supportData, minConf = 0.7):
    bigRuleList = []
    for i in range(1, len(L)): # 从二项集开始计算
    for freqSet in L[i]: # freqSet 为所有的 k 项集
    # 求该三项集的所有非空子集，1 项集，2 项集，直到 k-1 项集，用 H1 表示，为 list 类型 , 里面为 frozenset 类型，
    H1 = list(freqSet)
    all_subset = []
    getSubset(H1, all_subset) # 生成所有的子集
    calcConf(freqSet, all_subset, supportData, bigRuleList, minConf)
    return bigRuleList
    if __name__ == '__main__':
    dataSet = data_translation
    L, supportData = apriori(dataSet, minSupport = 0.02)
    rule = gen_rule(L, supportData, minConf = 0.35)
    frozenset({' 水果 / 蔬菜汁 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.02664 置信度：0.368495 lift 值为：1.44216
    frozenset({' 人造黄油 '}) --> frozenset({' 全脂牛奶 '}) 支
```

持度 0.024199 置信度：0.413194 lift 值为：1.617098

frozenset({' 仁果类水果 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.030097 置信度：0.397849 lift 值为：1.557043

frozenset({' 牛肉 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.021251 置信度：0.405039 lift 值为：1.58518

frozenset({' 冷冻蔬菜 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.020437 置信度：0.424947 lift 值为：1.663094

frozenset({' 本地蛋类 '}) --> frozenset({' 其他蔬菜 '}) 支持度 0.022267 置信度：0.350962 lift 值为：1.813824

frozenset({' 黄油 '}) --> frozenset({' 其他蔬菜 '}) 支持度 0.020031 置信度：0.361468 lift 值为：1.868122

frozenset({' 本地蛋类 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.029995 置信度：0.472756 lift 值为：1.850203

frozenset({' 黑面包 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.025216 置信度：0.388715 lift 值为：1.521293

frozenset({' 糕点 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.033249 置信度：0.373714 lift 值为：1.462587

frozenset({' 酸奶油 '}) --> frozenset({' 其他蔬菜 '}) 支持度 0.028876 置信度：0.402837 lift 值为：2.081924

frozenset({' 猪肉 '}) --> frozenset({' 其他蔬菜 '}) 支持度 0.021657 置信度：0.375661 lift 值为：1.941476

frozenset({' 酸奶油 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.032232 置信度：0.449645 lift 值为：1.759754

frozenset({' 猪肉 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.022166 置信度：0.38448 lift 值为：1.504719

frozenset({' 根茎类蔬菜 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.048907 置信度：0.448694 lift 值为：1.756031

frozenset({' 根茎类蔬菜 '}) --> frozenset({' 其他蔬菜 '}) 支持度 0.047382 置信度：0.434701 lift 值为：2.246605

frozenset({' 凝乳 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.026131 置信度：0.490458 lift 值为：1.919481

frozenset({' 热带水果 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.042298 置信度：0.403101 lift 值为：1.577595

frozenset({' 柑橘类水果 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.030503 置信度：0.36855 lift 值为：1.442377

frozenset({' 黄油 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.027555 置信度：0.497248 lift 值为：1.946053

frozenset({' 酸奶 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.056024 置信度：0.401603 lift 值为：1.571735

frozenset({' 其他蔬菜 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.074835 置信度：0.386758 lift 值为：1.513634

frozenset({' 酸奶 ', ' 全脂牛奶 '}) --> frozenset({' 其他蔬菜 '}) 支持度 0.022267 置信度：0.397459 lift 值为：2.054131

frozenset({' 酸奶 ', ' 其他蔬菜 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.022267 置信度：0.512881 lift 值为：2.007235

frozenset({' 全脂牛奶 ', ' 根茎类蔬菜 '}) --> frozenset({' 其他蔬菜 '}) 支持度 0.023183 置信度：0.474012 lift 值为：2.44977

frozenset({' 其他蔬菜 ', ' 根茎类蔬菜 '}) --> frozenset({' 全脂牛奶 '}) 支持度 0.023183 置信度：0.48927 lift 值为：1.914833

根据输出结果，对其中 4 条进行解释分析如下：

1. {' 其他蔬菜 '，' 酸奶 '}=>{' 全脂牛奶 '} 支持度约为 2.23%，置信度约为 51.29%。说明同时购买酸奶、其他蔬菜和全脂牛奶这 3 种商品的概率达 51.29%，而这种情况发生的可能性约为 2.23%。

2. {' 其他蔬菜 '}=>{' 全脂牛奶 '} 支持度最大约为 7.48%，置信度约为 38.68%。说明同时购买其他蔬菜和全脂牛奶这两种商品的概率达 38.68%，而这种情况发生的可能性约为 7.48%。

3. {' 根茎类蔬菜 '}=>{' 全脂牛奶 '} 支持度约为 4.89%，置信度约为 44.87%。说明同时购买根茎类蔬菜和全脂牛奶这 3 种商品的概率达 44.87%，而这种情况发生的可能性约为 4.89%。

4. {' 根茎类蔬菜 '}=>{' 其他蔬菜 '} 支持度约为 4.74%，置信度约为 43.47%。说明同时购买根茎类蔬菜和其他蔬菜这两种商品的概率达 43.47%，而这种情况发生的可能性约为 4.74%。

由上分析可知，顾客购买酸奶和其他蔬菜的时候会同时购买全脂牛奶，其置信度最大达到 51.29%。因此，顾客同时购买其他蔬菜、根茎类蔬菜和全脂牛奶的概率较高。

对于模型结果，从购物者角度进行分析：现代生活中，大多数购物者为“家庭煮妇”，购买的商品大部分是食品，随着生活质量的提高和健康意识的增加，其他蔬菜、根茎类蔬菜和全脂牛奶均为现代家庭每日饮食的所需品。因此，其他蔬菜、根茎类蔬菜和全脂牛奶同时购买的概率较高，符合人们的现代生活健康意识。

模型应用

模型结果表明：顾客购买其他商品的时候会同时购买全脂牛奶。因此，商场应该根据实际情况将全脂牛奶放在顾客购买商品的必经之路上，或是放在商场显眼的位置，以方便顾客拿取。顾客同时购买其他蔬菜、根茎类蔬菜、酸奶油、猪肉、黄油、本地蛋类和多种水果的概率较高，因此商场可以考虑捆绑销售，或者适当调整商场布置，将这些商品的距离尽量拉近，从而提升顾客的购物体验。

数据分析行业会员单位
——数据分析师事务所

编辑 / 数据委会员处 李苗苗 日期 / 2021-11

一、数据分析师事务所介绍：

数据分析师事务所（以下称“事务所”）是全新的第三方独立服务机构，是在工商局注册备案后成立，经中国商业联合会数据分析专业委员会（以下称“数据委”）审批通过成为数据分析行业会员，并授予中国数据分析行业会员执业资质证书，接受数据委的监督和管理。事务所统一的服务标准规范，是数据分析行业走向规范和自律的中流砥柱，是促进数据分析行业健康发展的中坚力量。

事务所由专业的数据分析师人才组成，可为企事业单位提供与大数据相关的专业服务。目前，我国数据分析师事务所会员遍布在全国各省，服务范围涉及数据分析相关领域，随着社会数字换转型的需求，广大企业对数据价值的认可，数据分析师事务所必将有更广阔的发展空间。

二、数据分析师事务所发展前景：

目前全国有百余家专业的数据分析师事务所遍布在各省市，业务基本围绕着数据的深度分析、业务场景构建、深层次的咨询，以帮助企业实现数字化转型的能力。近些年随着广大政府和企业对大数据的重视，数据会越来越多，技术门槛会越来越低，但是大量的数据进行深层次的分析就成为企业竞争的核心、成为企业大数据变现的核心，这对全国的数据分析师事务所来说是一个商业价值巨大的朝阳行业。

三、数据分析师事务所主要业务范围

事务所主要业务范围包括：数据分析咨询顾问服务、企业经营类数据分析服务、搭建大数据业务场景服务、综合解决方案服务、行业数据分析与解决方案服务等有关数据分析的业务等等。其中：

数据分析咨询顾问服务主要包括业务能力咨询、数据能力咨询、技术能力咨询、数据洞察能力咨询、解决方案咨询等咨询顾问服务等。

企业经营类数据分析服务包括财务分析、市场数据分析、客户数据分析、销售数据分析、质量数据分析、采购数据分析、人力资源数据分析等有关企业经营行为的数据分析等。

搭建大数据业务场景服务主要指模拟客户的业务场景，帮客户构建商业运营数据模型。

综合解决方案服务包括数据产品、数据平台建设等综合解决方案等。

四、创办数据分析师事务所可享受的数据委福利：

1、提供品牌宣传机会；

2、享受一年不少于 4 次实战案例学习，帮助事务所不断提升大数据专业能力；

3、享受一年一度执业教育免费培训；

4、帮助事务所解决数据分析师人才不足问题；

5、给予事务所资质在中国数据分析行业官网上备案、查询；

6、享受不定期举办免费且形式各异的创业指导活动，有效指导事务所的长期发展；

7、 优先参加数据委线上、线下具有一定影响力的活动；

8、 提供 Datahoop 智能大数据分析平台最高使用权限。

五、联系方式：

中国商业联合会数据分析专业委员会

网址：www.chinacpda.org

邮箱：xiehui@chinacpda.org

CPDA® 数据分析师

CERTIFIED PROJECTS DATA ANALYST. SINCE 2003

恒心

“为学需刚与恒
不刚则堕落
不恒则退”

数据分析 · 因你而不凡!

www.chinacpda.com | www.cpda.cn

TEL. 400-050-6600