尽管披着大数据的外衣
但这是一本常识书
每个人都可以从数据中获得知识和快乐

数据不说谎

大数据之下的世界

城市数据团◎编著

清华大学出版社

数据不说谎

大数据之下的世界

城市数据团◎编著

清华大学出版社
北 京

内容简介

这是一本让你"脑洞大开"的图书，让你尝试从大数据角度来解读这个世界，你会发现，有些问题，和你的直觉完全不一样。本书内容分为三部分：第一部分可概括为"脑洞大开"，以淘宝、旅游、餐馆取名等不同的角度切入，说明数据可以用于做许多有趣的事情；第二部分为数据与工作，包括公务员、二三线城市的衰落、创业等若干热门话题；第三部分为数据与生活，包括用数据帮助理解生活现象、用数据挖掘生活中的趣味，以及用数据看房市三个专题。

本书既适合大中专学生作为开阔眼界，拓展思维，帮助学习之用，也适合职场人士提升技能、辅助工作决策所用，是一切数据思维爱好者不可多得的好书。

图书在版编目(CIP)数据

数据不说谎：大数据之下的世界/城市数据团编著. —北京：清华大学出版社，2017
ISBN 978-7-302-46629-1

Ⅰ. ①数…　Ⅱ. ①城…　Ⅲ. ①数据处理　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 030992 号

责任编辑：刘志彬
封面设计：汉风唐韵
责任校对：宋玉莲
责任印制：杨　艳

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京亿浓世纪彩色印刷有限公司
经　　销：全国新华书店
开　　本：170mm×240mm　　**印　　张**：20　　**字　　数**：318 千字
版　　次：2017 年 6 月第 1 版　　**印　　次**：2017 年 6 月第1 次印刷
定　　价：69.00 元

产品编号：070835-01

前言　我们在用数据做什么

在这本书的最开始，我们想要提出这样一个问题：

谁最了解你？

是自己？

是配偶/恋人？

是父母/子女？

是同学/同事/朋友？

毫无疑问，以上几种人都存在于我们的生命中。

但是，跟“它”比起来，以上几种人对我们的了解恐怕都不够全面和客观。

没错，“它”就是手机，与我们形影不离的手机。

看看你手机上的那一大堆 APP——

微博和朋友圈知道，你今天心情好不好。

支付宝知道，你买了什么东西、花了多少钱。

微信和 QQ 知道，你都有哪些朋友，你跟哪些朋友的交流更密切。

豆瓣、知乎、每日头条知道，你都喜欢浏览哪些帖子和新闻。

虾米和酷狗知道，你喜欢听什么歌。

优酷和 B 站知道，你喜欢看什么视频。

饿了么和美团知道，你喜欢什么菜系和口味。

……

就算你什么 APP 也没装，只要你有一部手机，“它”就知道你什么时候工作，什么时候休息，知道你去了哪里，待了多久。

在手机面前，我们简直无所遁形。手机所知道的你，可能比你所知道的自己，更为真实。

而这些，都是我们自己告诉手机的。我们的每一次浏览、点赞、评论、下单、聊天，都以数据的形式被记录、被沉淀，最终塑造出了我们自己。

所以，请不要被“大数据”“开放数据”“数据挖掘”“深度学习”“神经网络”“云计算”“DMP”等奇奇怪怪的词汇所吓倒。我们每个人每天的生活起居、衣食住行，都在产生数据，并享受着数据给我们带来的便利服务。

事实上，数据已经和我们的视觉、听觉、触觉一样，成为了帮助我们去了解自己、了解他人、了解事物的重要方法。

与其他信息源相比，数据更有可能提供全面和客观的信息，从而帮助我们更快速和高效地了解问题、解决问题。

例如，你母亲催你去相亲，并提供了 100 位相亲者的资料。显然，你不可能一个个把他们约出来见面，一个个去了解和评价他们——你甚至都不可能仔细读完这 100 份资料。

我们通常的做法是，设立一些限制条件，对年龄、身高、学历、收入等进行筛选，再逐份阅读符合条件的相亲者的资料，直到将相亲对象数量减少到个位数。如此，我们的相亲效率就大大提高了。

然而，在享受数据给我们带来的高效便利的同时，我们还必须意识到：数据分析只能提供结果，不能提供结论；数据之所以能做许多事情，是因为使用数据的人做了很多的思考。

例如，2013 年，Amazon Studios 和 Netflix，美国的两家传媒公司，都对自己网站上客户的视频浏览行为进行了分析。接受分析的浏览行为包括客户看了什么视频、什么时候看的、在何处暂停、在何处跳过、在何处反复观看、给视频的评分等。

根据数据分析的结果，两家公司一致认为观众会对政治主题感兴趣，但在视频的体裁、制作等方面则有着完全不同的认知。而后，Amazon Studios 推出了由四位议员作为主角的情景喜剧，Netflix 则推出只有一位议员作为主角的电视连续剧。前一部作品名为《阿尔法屋》(*Alpha House*)，观众反应平平；后一

部作品则是风靡一时、获奖无数的《纸牌屋》(*House of Cards*)。

所以，即使在一个“大数据”炙手可热、喧嚣尘上的时代，人仍然是主体，是人的智慧让数据具有了价值。

我们，城市数据团的小伙伴们，就是这样一群人：利用数据去了解城市的发展、挖掘城市生活中有趣的故事。对我们而言，数据是帮助我们认识城市的工具、帮助我们在城市里更好地生活的工具，而通过数据发现的东西才是价值和乐趣所在。我们乐意将这些发现拿出来共享。

本书由城市数据团组织编写，并写作了本书的大部分章节。城市数据团的主要成员包括高路拓、汤舸、王咏笑、王宇鹏等。参与了本书部分章节写作的其他数据团成员和合作伙伴包括(按文章收录顺序)：

陈宇佳(1.1.2)、郭斌亮(1.2.1)、陈至奕(1.2.3/2.1.1)、冯里婧(2.1.2)、钱骏杰(3.1.2)、张慈(3.1.3)、曹新(3.1.5)、曹湛(3.2.4)、韩旭(3.2.5)、方娴(3.3.1)、张健(3.3.2)、衣霄翔(3.3.3)、陈晨(3.3.4)。

除写作团队之外，感谢以下机构对本书内容提供了数据支持和技术支持(按文章收录顺序)：

- 银联智惠信息服务(上海)有限公司(1.1.1/3.1.1/3.2.3)
- 滴滴大数据研究中心(1.2.2/3.2.2)
- 小猿搜题(2.1.1)
- BDP 个人版(2.1.2/3.1.4)
- TalkingData(2.1.4/3.2.2/3.2.4)
- 阿里研究院(2.2.1)
- 大众点评研究院(2.2.2/3.1.2/3.1.3/3.2.5)
- 上海道融自然保护与可持续发展中心(3.1.5)
- 同策房产咨询(3.3.1)

本书由城市数据团这个活跃在互联网上的大数据团队完成。如果您看完本书以后，能够增加一些对这个数据时代的了解、愿意去热爱数据和使用数据，将是对我们莫大的鼓励。

城市数据团

2017 年 3 月

目 录

第 1 章

数据，另一种视角

你消费吗？旅游吗？上班吗？

你知道别人是怎么消费、怎么旅游、怎么上班的吗？

我们对于世界和城市的认知，往往来源于自己和身边其他人的生活经验。

所以，我们的认知往往是主观化和碎片化的。

但是，当我们拥有了“数据”这个工具的时候，我们就获得了重新认识世界的机会。

1.1 数据之下的中国

本节内容主要涉及一个主题：如何脑洞大开地搜集和利用各种数据，以非常规的方式呈现出中国经济发展的三个截面。

数据之下的中国，是一个让你既熟悉又新鲜的中国。

1.1.1 2015年，中国人是怎么花钱的

在一波接一波的寒潮侵袭之后，期盼已久的春节假期终于到了。

同事同学们纷纷放假回家，连亲爱的学姐也不在上海，只留我一个人凄冷地坐在工作台前，独自迎接假期前最难熬的几天。

一个人的时候，总是会想很多。

是的。回首即将逝去的羊年，我感慨万千。虽然不出意外地又（为什么要加一个又字呢）穷困潦倒地度过了漫长的一年，但幸运的是在这期间认识了不少天南海北的朋友。

因此，虽然还在孤独地加班，但我仍然心系着祖国人民，安静地准备完成春节前的最后一项数据工作：

年度全国消费数据总盘点。

好吧，问题来了——

Q1：2015 年，全国人民到底花了多少钱？

2015 年全球范围内可使用银联卡商户共 3 390 万家，ATM 共 200 万台，境外共发行银联卡 5 200 万张。

根据刷卡交易统计，2015 年全年，全国人民的刷卡交易总金额达到 53.9 万亿元。

53.9 万亿元，是个什么概念呢？

我们可以想象一下：如果把这 53.9 万亿元全换成 100 元的人民币钞票，并将其一张一张紧挨着排列起来的话，这些钱大概可以绕地球赤道 2 100 圈；从地球排到太阳的话，可以走一半多一点的路程。

假如这还想象不出来的话，我们可以换个角度来看：

根据国家统计局的数据，2015 年，全国 GDP 总额约为 67.7 万亿元。也就是说，仅是刷卡消费，全国人民就刷掉了年度国内生产总值的 79.6%。

亲爱的，你 2015 年创造了多少 GDP？又刷掉了多少份额呢？

算好了吗？

好的话，我们不妨再来研究一下第二个问题，看看你的消费和全国总体水平相比如何呢？问题来了。

Q2：这 53.9 万亿元，都是怎么花掉的呢？

首先，让我们来看看这些钱是在什么时间内被花掉的呢？

我们统计了境内日均刷卡的交易金额，并将其细分到每一个小时。2015 年日均逐小时交易曲线见图 1-1，银联卡交易类型占比见图 1-2。

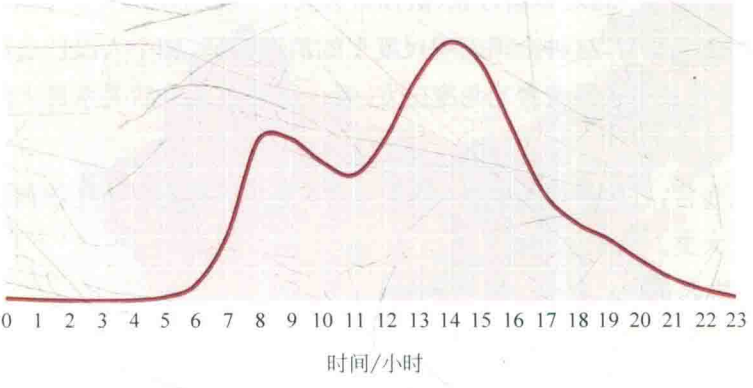

图 1-1　日均逐小时交易曲线

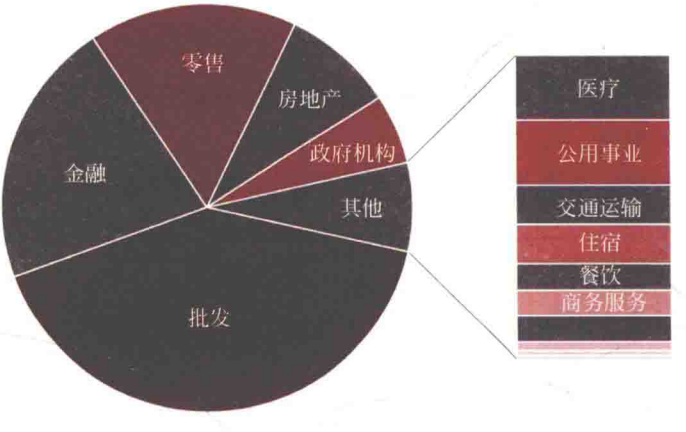

图 1-2　2015 年银联卡交易类型占比

假如我们把 2015 年全年浓缩到一天来看的话，可以发现：14:00～15:00 和 8:00～9:00 是全国人民刷卡的高峰时段，分别占全天交易总额的 12% 与 8%。

亲爱的，你的卡是不是在这个时段内被刷爆的呢？

看完了交易时间，我们再来看一下交易的类型。我们将年度刷卡交易总金额分配到交易类型上，如下所述。

(1) 从全国尺度上来看，最多的刷卡交易金额发生在批发行业，份额第

一，大概可以购买 16 个阿里巴巴。

（2）份额第二的是金融行业，大概可以购买 7 个中国工商银行。

（3）份额第三的是零售行业（俗称买买买），大概可以购买 5 个沃尔玛。

也许你会觉得，这种全国宏观尺度上的消费特征，和个人没什么关系。那么，我们不妨从个人消费者的角度出发，看一下与市民生活关系最大的消费门类吧。

一般而言，各种消费类型中，与市民生活关系最大的应该是衣食住行金融教育六个大类。结果如何呢？

（1）排名第一：金融。毫无悬念。

（2）排名第二：住房。其交易总额大约是金融类的三分之一。

（3）排名第三：旅游。虽然交易总额排名第三，但也不过是住房类的零头而已。

（4）排名第四：衣（衣物类零售）。其总额大约是旅游的三分之一。

（5）排名第五：吃（餐饮）。交易总额与衣物类零售不相伯仲。

（6）排名最后：教育。其交易总额大约是餐饮的 70%。没错，这个结果毫不意外、发人深省。

亲爱的，你的消费结构和全国人民相比，究竟怎样呢？

每个人的消费结构自然千奇百怪。

且不说个人，即使从省市的角度上去区分，也可以看到消费结构上的巨大差异。我们来看看：

Q3：全国各省的消费结构有什么样的偏好呢？

我们仍然将数据聚焦在衣物、餐饮、住房、旅游、金融和教育六个大类上。然后将各类消费金额占总消费金额的比例作为消费偏好的核心指标，分配到各省，可以得到以下结果。

（1）**衣物类**消费偏好前五名省市：云南、浙江、甘肃、山西、湖北。

想必云南四季如春，民族众多，姑娘们想怎么打扮就怎么打扮吧。见图 1-3。

（2）**餐饮类**消费偏好前五名省市：海南、上海、西藏、宁夏、北京。

吃货集聚在上海、北京，这点毫不意外。但没想到海南、西藏、宁夏等边远

地区的吃货能量同样惊人，见图 1-4。

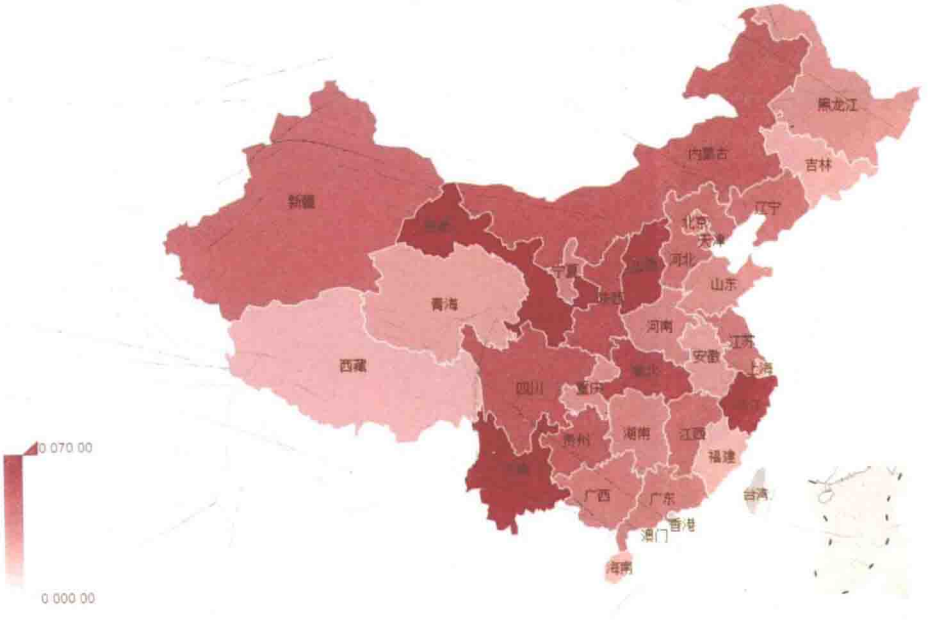

图 1-3　各省衣物类消费占比

图 1-4　各省餐饮类消费占比

（3）住房类消费偏好前五名省市：海南、四川、贵州、北京、安徽。

非常出乎意料的，前三名竟然不是以高房价著称的北上广哦！看来虽然北上广的绝对房价居高不下，但从真实的消费结构上，海南和四川的房价水平也不容小觑。相比北京排名第四，而上海甚至都没有挤进前五，见图 1-5。

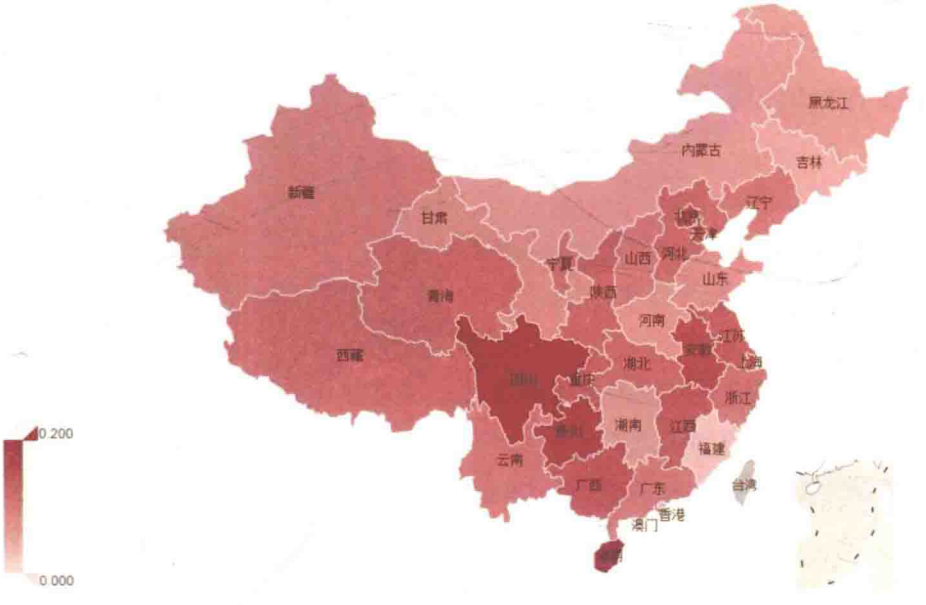

图 1-5　各省住房类消费占比

（4）旅游类消费偏好前五名省市：西藏、海南、青海、新疆、云南。

从图 1-6 可以看到，西部的旅游消费偏好明显高于东部。而排名前五的省市，也都是以旅游胜地著称的地区。

（5）金融类消费偏好前五名省市：福建、重庆、广东、湖南、上海。

从图 1-7 可以看到，我国东南地区在金融类消费偏好中可谓一枝独秀，福建省拔得头筹。排名前五的省市中，上海市已经是最北方的地区了。

（6）教育类消费偏好前五名省市：陕西、四川、北京、海南、湖南。

从图 1-8 可以看到，陕西省、四川省在教育类消费上的偏好明显高于全国其他地区。我在想，这些地方的孩子们是不是从幼儿园就开始上补习班了？

说明一下：本书消费数据中没有统计到中国台湾地区的数据，所以地图上台湾地区的颜色与其他省市不同。

图 1-6　各省旅游类消费占比

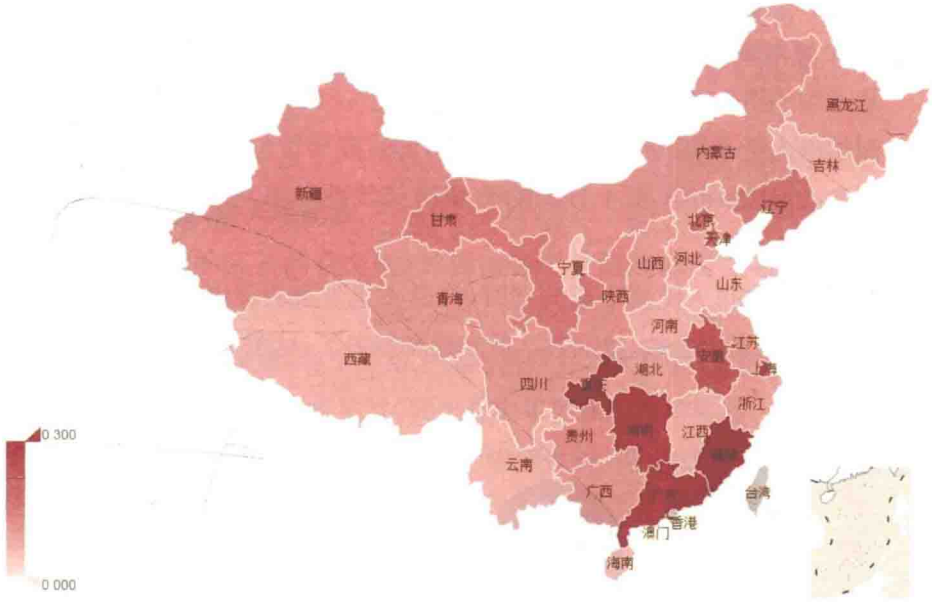

图 1-7　各省金融类消费占比