Atitit 数据分析之道 attilax著

[1. 数据分析的目的 2](#_Toc24089)

[1.1. 描述型：发生了什么？可视化仪表盘化很重要。 3](#_Toc25394)

[1.2. 2. 诊断型：为什么会发生？结合历史记录 3](#_Toc4426)

[1.3. 3. 预测型：可能发生什么？ 4](#_Toc10190)

[1.4. 4. 指导型 规划未来行动：我需要做什么？ 4](#_Toc14094)

[2. 数据分析 4](#_Toc12628)

[3. 与数据挖掘的区别 5](#_Toc10523)

[3.1. 数据挖掘更侧重于机器对未来的预测，一般应用于分类、聚类、推荐、关联规则等。 5](#_Toc12783)

[3.2. 与数据分析的区别 5](#_Toc6957)

[4. 数据类型 .1  数据库、SQL和大数据 2 6](#_Toc6069)

[5. 数据概念 6](#_Toc28424)

[5.1. 1.2  绘制数据结构 4 6](#_Toc24893)

[5.2. 1.3  使用数据流描述数据分析12 7](#_Toc17728)

[5.3. 数据的类型 17 2.2.1　总体数据和样本数据 17 2.2.2　数量数据和属性数据 17 2.2.3　截面数据和时间序列数据 17 2.2.4　数据的来源 18 7](#_Toc16318)

[5.4. 2.4　数据的分布 24 2.4.1　属性数据的频数分布 24 2.4.2　频率分布 26 2.4.3　数量数据的频率分布 26 2.4.4　直方图 29 2.4.5　累积分布 32 8](#_Toc6495)

[6. 第2章　描述统计分析 15 8](#_Toc15135)

[7. 常见数据分析方法 8](#_Toc27686)

[7.1. 3 最优化：寻找最大值 4 数据图形化：图形让你更精明 5 假设检验：假设并非如此 6 贝叶斯统计：穿越第一关 7 主观概率：信念数字化 8 启发法：凭人类的天性作分析 9 直方图：数字的形状 10 回归：预测 11 误差：合理误差 12 相关数据库：你能关联吗？ 13 整理数据：井然有序 9](#_Toc28587)

[8. 数据探索方法 9](#_Toc21760)

[8.1. 2.2 Excel中的绘图 40 三大图柱形图饼图 曲线图 9](#_Toc31301)

[8.2. 2.4.1 直方图 55 10](#_Toc23518)

[8.3. 2.5  探索更多的值——最小值、最大值和模式 64 10](#_Toc3914)

[8.4. 2.8  由一个列的数据扩展到所有列的数据汇总 73 10](#_Toc18193)

[9. 第3章　数据可视化 66 11](#_Toc22923)

[10. 3.1  基本的统计学概念 80 12](#_Toc9441)

[11. 第4章  发生的地点在何处？ 119 12](#_Toc27135)

[11.1. 4.1  纬度和经度 120 13](#_Toc22258)

[11.2. 4.3  地理等级 142 13](#_Toc1568)

[12. 第5章  关于时间 161 14](#_Toc21941)

[12.1. 5.3  两个日期之间有多长？178 14](#_Toc21004)

[12.2. 5.4  跨年比较 188 15](#_Toc21619)

[13. 第6章  客户的持续时间有多久？使用生存分析理解客户和他们的 16](#_Toc31975)

[13.1. 6.4  对比不同的客户分组 230 17](#_Toc29375)

[14. 第10章  关联规则 387 22](#_Toc2319)

[15. 第11章  SQL数据挖掘模型 423 23](#_Toc20037)

[16. 第12章  最佳拟合线：线性回归 25](#_Toc22301)

[17. 第14章  性能问题：高效使用SQL543 28](#_Toc8281)

[18. 其他 31](#_Toc4556)

[19. 第1讲 大数据简介 3 32](#_Toc11439)

[20. 线性回归分析 118 33](#_Toc19005)

[21. 时间序列分析与预测 182 34](#_Toc13092)

[22. 第6章　数据挖掘 226 35](#_Toc20269)

[23. 第12章　决策分析 477 36](#_Toc25730)

[24. 参考资料 37](#_Toc24352)

[24.1. 《深入浅出数据分析》(（美）米尔顿　著)【简介\_书评\_在线阅读】 - 当当图书.mhtml 37](#_Toc6815)

[24.2. 《数据分析技术(第2版) 使用SQL和Excel工具》([美]Gordon S. Linoff 著 陶佰明 译)【简介\_书评\_在线阅读】 - 当当图书.mhtml 37](#_Toc5255)

[24.3. 《大数据分析师权威教程 大数据分析与预测建模》(Wrox国际IT认证项目组)【简介\_书评\_在线阅读】 - 当当图书.mhtml 38](#_Toc22095)

[24.4. 《商业数据分析》([美]杰弗里 D.坎姆（Jeffrey D.Camm）等)【简介\_书评\_在线阅读】 - 当当图书.mhtml 39](#_Toc9321)

# 数据分析的目的

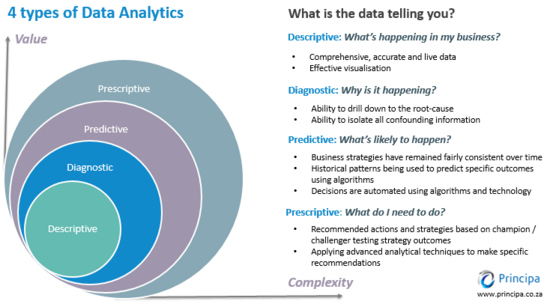
[描述型：发生了什么？可视化仪表盘化很重要。 1](#_Toc16110)

[2. 诊断型：为什么会发生？结合历史记录 1](#_Toc12172)

[3. 预测型：可能发生什么？ 2](#_Toc31000)

[4. 指导型 规划未来行动：我需要做什么？ 2](#_Toc15793)

[数据分析 2](#_Toc29147)



在这篇博文中，我们聚焦于在数据科学领域所遇到的四种类型的数据分析模式：描述型、诊断型、预测型和指导型。

简单地来说，分析可被划分为4个重要的类别。下面我将详细解释这四个类别。

## 描述型：发生了什么？可视化仪表盘化很重要。

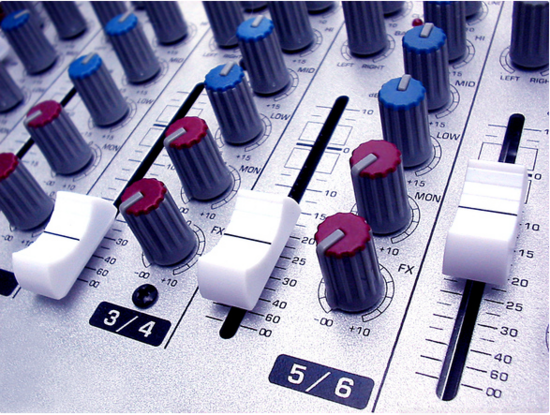
## 2. 诊断型：为什么会发生？结合历史记录

设计良好的商业信息dashboard整合了时间序列数据（譬如，在多个联系时间点上的数据）的读入、特征的过滤和钻入功能，能够用于这类分析。

Aop接入日志分析。

## 3. 预测型：可能发生什么？

## 4. 指导型 规划未来行动：我需要做什么？



在价值和复杂度上，下一步就是指导性模型。指导性模型基于发生了什么、为什么会发生以及一系列“可能发生什么”的分析，帮助用户确定要采取的最好的措施。很显然，指导性分析不是一个单独的行为，实际上它是其他很多行为的主导。

交通应用是一个很好的例子，它帮助人们选择最好的回家路线，考虑到了每条路线的距离、在每条路上的速度、以及很关键的目前的交通限制。

# 数据分析

四种类型的数据分析模式 - 推酷.mhtml

# 与数据挖掘的区别

## 数据挖掘更侧重于机器对未来的预测，一般应用于分类、聚类、推荐、[关联规则](https://www.baidu.com/s?wd=%E5%85%B3%E8%81%94%E8%A7%84%E5%88%99&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1Ykuj64mWKbuHnsn1R4uHmv0ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3EPjbvrjfdPj0drjc3PHb1PHcY" \t "https://zhidao.baidu.com/question/_blank)等。

从分析的目的来看，数据分析一般是对历史数据进行统计学上的一些分析，数据挖掘更侧重于机器对未来的预测，一般应用于分类、聚类、推荐、[关联规则](https://www.baidu.com/s?wd=%E5%85%B3%E8%81%94%E8%A7%84%E5%88%99&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1Ykuj64mWKbuHnsn1R4uHmv0ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3EPjbvrjfdPj0drjc3PHb1PHcY" \t "https://zhidao.baidu.com/question/_blank)等。

数据挖掘（Data mining）是一个跨学科的计算机科学分支。它是用人工智能、机器学习、统计学和数据库的交叉方法在相对较大型的数据集中发现模式的计算过程。 数据挖掘过程的总体目标是从一个数据集中提取信息，并将其转换成可理解的结构，以进一步使用。除了原始分析步骤，它还涉及到数据库和数据管理方面、数据预处理、模型与推断方面考量、兴趣度度量、复杂度的考虑，以及发现结构、可视化及在线更新等后处理。数据挖掘是“数据库知识发现”（KDD）的分析步骤

作者：周圆方  
链接：https://www.zhihu.com/question/20127962/answer/150275661  
来源：知乎  
著作权归作者所有。商业转载请联系作者获得授权，非商业转载请注明出处。

## 与数据分析的区别

从分析的目的来看，数据分析一般是对历史数据进行统计学上的一些分析，数据挖掘更侧重于机器对未来的预测，一般应用于分类、聚类、推荐、[关联规则](https://www.baidu.com/s?wd=%E5%85%B3%E8%81%94%E8%A7%84%E5%88%99&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1Ykuj64mWKbuHnsn1R4uHmv0ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3EPjbvrjfdPj0drjc3PHb1PHcY" \t "https://zhidao.baidu.com/question/_blank)等。  
从分析的过程来看，数据分析更侧重于统计学上面的一些方法，经过人的推理演译得到结论；数据挖掘更侧重由机器进行自学习，直接到得到结论。  
从分析的结果看，数据分析的结果是准确的统计量，而数据挖掘得到的一般是模糊的结果。

“数据分析”的重点是观察数据，“数据挖掘”的重点是从数据中发现“知识规则”KDD（Knowledge Discover in Database）。  
  
“数据分析、数据统计”得出的结论是人的智力活动结果，“数据挖掘”得出的结论是机器从学习集（或训练集、样本集）发现的知识规则。  
“数据分析”需要人工建模，“数据挖掘”自动完成[数学建模](https://www.baidu.com/s?wd=%E6%95%B0%E5%AD%A6%E5%BB%BA%E6%A8%A1&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1Ykuj64mWKbuHnsn1R4uHmv0ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3EPjbvrjfdPj0drjc3PHb1PHcY" \t "https://zhidao.baidu.com/question/_blank)。

# 数据类型 .1  数据库、SQL和大数据 2

1.1.1 什么是大数据？ 2

1.1.2 关系型数据库 3

1.1.3 Hadoop和Hive 3

1.1.4 NoSQL和其他类型的数据库 3

1.1.5 SQL 4

# 数据概念

## 1.2  绘制数据结构 4

1.2.1 什么是数据模型？ 5

1.2.2 什么是表？ 5

1.2.3 什么是实体-关系图表？ 8

1.2.4 邮政编码表 9

1.2.5 订阅数据集 10

1.2.6 订单数据集 11

1.2.7 关于命名的提示 12

## 1.3  使用数据流描述数据分析12

1.3.1 什么是数据流？ 13

1.3.2 数据流、SQL和关系代数 16

## 数据的类型 17 2.2.1　总体数据和样本数据 17 2.2.2　数量数据和属性数据 17 2.2.3　截面数据和时间序列数据 17 2.2.4　数据的来源 18

## 2.4　数据的分布 24 2.4.1　属性数据的频数分布 24 2.4.2　频率分布 26 2.4.3　数量数据的频率分布 26 2.4.4　直方图 29 2.4.5　累积分布 32

# 第2章　描述统计分析 15

# 常见数据分析方法

## 3 最优化：寻找最大值 4 数据图形化：图形让你更精明 5 假设检验：假设并非如此 6 贝叶斯统计：穿越第一关 7 主观概率：信念数字化 8 启发法：凭人类的天性作分析 9 直方图：数字的形状 10 回归：预测 11 误差：合理误差 12 相关数据库：你能关联吗？ 13 整理数据：井然有序

# 数据探索方法

## 2.2 Excel中的绘图 40 三大图柱形图饼图 曲线图

2.2.1 基础图表：柱形图 41

2.2.2 单元格中的条形图 45

2.2.3 柱形图的有用变化形式 47

2.2.4 其他类型的图表 50

2.3  迷你图 53

2.4  列中包含的值 55

## 2.4.1 直方图 55

2.4.2 计数的直方图 58

2.4.3 计数的累积直方图 60

2.4.4 数字值的直方图(频率) 60

## 2.5  探索更多的值——最小值、最大值和模式 64

2.5.1 最小值和最大值 64

2.5.2 最常见的值(模式) 65

2.6  探索字符串值 66

2.6.1 长度的直方图 66

2.6.2 起始或结尾包含空白字符

的字符串 66

## 2.8  由一个列的数据扩展到所有列的数据汇总 73

2.8.1 针对单列的汇总 74

2.8.2 返回表中所有列的查询 76

2.8.3 使用SQL生成汇总编码 76

# 第3章　数据可视化 66

3.1　概述 68  
3.2　表格 70  
3.2.1　表格设计原则 71  
3.2.2　交叉表 73  
3.2.3　Excel数据透视表 75  
3.3　图 79  
3.3.1　散点图 79  
3.3.2　折线图 81  
3.3.3　条形图和柱状图 83  
3.3.4　饼状图和3D图的评述 86  
3.3.5　气泡图 86  
3.3.6　热点图 88  
3.3.7　其他多变量图形 90  
3.3.8　Excel中的数据透视图 94  
3.4　高级可视化方法 96  
3.4.1　高级图形 96  
3.4.2　地理信息系统图 98  
3.5　数据仪表盘 99  
3.5.1　制作数据仪表盘的原则 99  
3.5.2　数据仪表盘的应用 99  
本章小结 101

# 3.1  基本的统计学概念 80

3.1.1 虚拟假设 80

3.1.2 可信度和概率 81

3.1.3 正态分布 82

3.2  平均值的区别有多大？ 85

3.2.1 方法 85

3.2.2 子集平均值的标准差 85

3.2.3 三个方法 87

3.3  对表做抽样 89

3.3.1 随机抽样 89

3.3.2 可重复的随机样本 90

3.3.3 分层比例抽样 91

3.3.4 平衡的样本 92

3.4  计数的可能性 93

3.4.1 有多少男性成员？ 96

# 第4章  发生的地点在何处？ 119

## 4.1  纬度和经度 120

4.1.1 纬度和经度的定义 120

4.1.2 度数、分钟和秒 121

4.1.3 两个位置之间的距离 122

4.1.4 包含邮政编码的图片 128

4.2  人口统计 131

4.2.1 极端情况：最富有的和最贫

穷的人 132

4.2.2 分别在使用订单和不使用订

单的情况下比较邮政编码 137

## 4.3  地理等级 142

4.3.1 州中最富有的邮政编码 142

4.3.2 州中拥有最多订单的邮政

编码 143

4.3.3 地理数据中有趣的层级

结构 145

4.3.4 计算郡的财富 148

4.3.5 财富值的分布 150

4.3.6 在郡中，哪个邮政编码是相对

最富有的？ 151

4.3.7 拥有最高的相对订单占有

份额的郡 152

4.4  在Excel中绘制地图 155

4.4.1 为什么绘制地图？ 155

4.4.2 不能绘图 156

4.4.3 网络地图 156

4.4.4 邮政编码散点图之上的州

边界 157

4.5  小结 159

# 第5章  关于时间 161

5.1  数据库中的日期和时间162

5.2  开始调研日期 166

5.2.1 确认日期中没有时间 166

5.2.2 根据日期比较计数 167

5.2.3 订单数和订单大小 172

5.2.4 星期 175

## 5.3  两个日期之间有多长？178

5.3.1 以天为单位的持续时间 178

5.3.2 以星期为单位的持续时间 180

5.3.3 以月为单位的持续时间 180

5.3.4 有多少个星期一？ 181

5.3.5 下一个周年纪念日(或生日)

是什么时候？ 184

## 5.4  跨年比较 188

5.4.1 以天为单位比较 188

5.4.2 以星期为单位比较 189

5.4.3 以月为单位比较 190

5.5  以天计算活跃客户数量196

5.5.1 某天的活跃客户数量 196

5.5.2 每天的活跃客户数量 196

5.5.3 有多少不同类型的客户？ 198

5.5.4 不同任期时段的客户数量 198

5.5.5 只使用SQL计算活跃客户 201

5.6 Excel中的简单图表动画 203

5.6.1 从订单生成日期到运货

日期 203

5.6.2 订单延时在每年中的变化 205

5.7  小结 208

# 第6章  客户的持续时间有多久？使用生存分析理解客户和他们的

价值 209

6.1  生存分析 210

6.1.1 平均寿命 211

6.1.2 医学研究 212

6.1.3 关于风险率的示例 212

6.2  风险计算 213

6.2.1 数据调研 214

6.2.2 风险率 216

6.2.3 客户可视化：时间与任期 217

6.2.4 截尾 219

6.3  生存率和保留率 220

6.3.1 生存率的点的估计 220

6.3.2 计算任意任期的生存率 221

6.3.3 在SQL中计算生存率 222

6.3.4 简单的客户保留率计算 225

6.3.5 保留率和生存率的区别 226

6.3.6 风险率和生存率的简单

示例 227

## 6.4  对比不同的客户分组 230

6.4.1 市场总结 230

6.4.2 市场分层 231

6.4.3 生存率比例 234

6.4.4 条件生存率 234

6.5  随时间变化的生存率 236

6.5.1 特定风险率随时间的变化 236

6.5.2 按照起始年份分类的客户

生存率 238

6.5.3 之前的生存率什么样？ 239

6.6  由生存率衍生出来的重要

指标 241

6.6.1 估算生存点 241

6.6.2 客户任期的中间值 242

6.6.3 客户生命周期的中间值 242

6.6.4 风险率的置信度 243

6.7  使用生存率计算客户价值245

6.7.1 估算收入 246

6.7.2 对个体的未来收入的估算 247

6.7.3 当前客户分组的收入估算 249

6.7.4 所有客户未来收入的估算 251

6.8  预测 253

6.8.1 对已有客户的预测 254

6.8.2 对新开始者的预测 258

6.9  小结 259

第7章  影响生存率的因素：客户

任期 261

7.1  哪些因素是重要的，何时

重要？ 262

7.1.1 方法说明 262

7.1.2 使用平均值比较数字因素 264

7.1.3 风险比例 268

7.2  左截断 271

7.2.1 认识左截断 271

7.2.2 左截断的影响 273

7.2.3 如何从理论上解决左截断

问题 274

7.2.4 估算一个任期的风险率 275

7.2.5 估算所有任期的风险率 276

7.2.6 在SQL中计算 277

7.3  时间窗 278

7.3.1 一个商业问题 278

7.3.2 时间窗=左截断 右截尾 278

7.4  竞争风险 283

7.4.1 竞争风险的示例 283

7.4.2 竞争风险的“风险率” 284

7.4.3 竞争风险的“生存率” 286

7.4.4 随着时间的变化，客户身上

发生了什么？ 287

7.5  事件前后 291

7.5.1 三种情况 291

7.5.2 使用生存率预测来理解一次

性事件 293

7.5.3 比较前后风险率 294

7.5.4 基于对列的方法 294

7.5.5 基于对列的方法：完全队列 295

7.5.6 事件影响的直接估计 297

7.6  小结 301

第8章  多次购买以及其他重复事件303

8.1  标识客户 304

8.1.1 谁是那个客户？ 304

8.1.2 其他客户信息 313

8.1.3 每一年出现多少新客户？ 316

8.2 RFM分析 325

8.2.1 维度 325

8.2.2 计算RFM单元格 329

8.2.3 RFM的有用程度 330

8.3  随着时间的变化，哪些家庭的

购买金额在增长？ 334

8.3.1 最早值和最晚值的比较 334

8.3.2 第一年和最后一年的值的

比较 341

8.3.3 最佳拟合线的趋势 343

8.4  距离下一次事件的时间344

8.4.1 计算背后的想法 344

8.4.2 使用SQL计算下一次购买

日期 345

8.4.3 从下一次购买日期到时间至

事件的分析 346

8.4.4 时间到事件分析的分层 347

8.5  小结 347

第9章  购物车里有什么？购物车

分析 349

9.1  探索产品 349

9.1.1 产品的散点图 350

9.1.2 产品组的运输年份 351

9.1.3 订单中的重复产品 353

9.1.4 单位数量的直方图 358

9.1.5 在一个订单中，哪个产品可能

出现多次购买的情况？ 359

9.1.6 改变价格 361

9.2  产品和客户价值 362

9.2.1 订单大小的一致性 362

9.2.2 与一次性客户关联的产品 365

9.2.3 与最好的客户相关的产品 368

9.2.4 剩余价值 370

9.3  产品的地理分布 372

9.3.1 每一个州中最常见的产品 372

9.3.2 哪些产品广受欢迎，哪些产品

只在本地受欢迎？ 373

9.4  哪些客户购买了指定产品？375

9.4.1 哪些客户拥有最受欢迎的

产品？ 375

9.4.2 客户拥有哪个产品？ 376

9.4.3 哪些客户有3个特定的

产品？ 381

9.4.4 普遍的嵌套集合的查询 384

9.5  小结 385

# 第10章  关联规则 387

10.1 项集 388

10.1.1 两个产品的组合 388

10.1.2 更常见的项集 391

10.1.3 家庭，而不是订单 396

10.2 最简单的关联规则 399

10.2.1 关联和规则 400

10.2.2 零项关联规则 400

10.2.3 概率的分布情况 401

10.2.4 零项关联告诉了我们

什么？ 402

10.3 单项关联规则 402

10.3.1 单项关联规则的价值 402

10.3.2 生成所有的单项规则 404

10.3.3 包含评估信息的单项

规则 405

10.3.4 基于产品组的单项规则 406

10.4 双项关联 407

10.4.1 计算双项关联 408

10.4.2 使用卡方找到最佳规则 409

10.4.3 异质相关 413

10.5 扩展关联规则 416

10.5.1 多项关联 416

10.5.2 一个查询中的多项关联 418

10.5.3 使用产品属性的规则 418

10.5.4 左右两侧项集内容不同

的规则 419

10.5.5 之前和之后：有序关联

规则 419

10.6 小结 422

# 第11章  SQL数据挖掘模型 423

11.1 定向数据挖掘介绍 424

11.1.1 定向模型 424

11.1.2 建模中的数据 425

11.1.3 建模应用示例 427

11.1.4 模型评估 429

11.2 相似性模型 429

11.2.1 模型是什么？ 430

11.2.2 最好的邮政编码是

哪个？ 430

11.2.3 基础的相似性模型 431

11.2.4 使用Z分数计算相似性

模型 433

11.2.5 邻近模型示例 434

11.3 最受欢迎产品的查找模型 435

11.3.1 最受欢迎的产品 435

11.3.2 计算最受欢迎的产品组 436

11.3.3 评估查找模型 437

11.3.4 使用调试查找模型做

预测 437

11.3.5 使用二元分类 439

11.4 用于订单大小的查找模型 440

11.4.1 最基本的模型：无维度

模型 440

11.4.2 添加一个维度 441

11.4.3 添加额外的维度 443

11.4.4 检查不稳定性 443

11.4.5 使用平均值图表评估

模型 444

11.5 用于响应率的查找模型 445

11.5.1 将整体概率作为一个

模型 445

11.5.2 探索不同的维度 446

11.5.3 模型的精准度 447

11.5.4 ROC图表和AUC 450

11.5.5 加入更多的维度 453

11.6 朴素贝叶斯模型(证据模型) 455

11.6.1 概率的一些概念 455

11.6.2 计算朴素贝叶斯模型 457

11.6.3 朴素贝叶斯模型：评分和

提升度 463

11.6.4 朴素贝叶斯模型和查找

模型的比较 465

11.7 小结 466

# 第12章  最佳拟合线：线性回归

模型 467

12.1 最佳拟合线 468

12.1.1 任期和支付金额 468

12.1.2 最佳拟合线的属性 469

12.1.3 小心数据 473

12.1.4 图表中的趋势线 474

12.1.5 使用LINEST()函数的

最佳拟合 479

12.2 使用R2衡量拟合程度 483

12.2.1 R2值 483

12.2.2 R2的局限性 484

12.2.3 R2的含义 484

12.3 直接计算最佳拟合线系数 485

12.3.1 计算系数 485

12.3.2 在SQL中计算最佳

拟合线 486

12.3.3 价格弹性 487

12.4 加权的线性回归 492

12.4.1 在第一年停止的客户 492

12.4.2 加权的最佳拟合 493

12.4.3 图表中的加权最佳

拟合线 494

12.4.4 SQL中的加权最佳

拟合线 495

12.4.5 使用Solver的加权最佳

拟合线 496

12.5 多个输入 498

12.5.1 Excel中的多维回归 498

12.5.2 建立包含三个变量的

模型 500

12.5.3 使用Solver处理多维

回归 501

12.5.4 逐个选择输入变量 501

12.5.5 SQL中的多维回归 502

12.6 小结 503

第13章  为进一步分析数据创建客户

签名 505

13.1 什么是客户签名？ 506

13.1.1 什么是客户？ 506

13.1.2 客户签名的源数据 507

13.1.3 使用客户签名 510

13.2 设计客户签名 511

13.2.1 调试和预测 511

13.2.2 字段的角色 511

13.2.3 时间段 512

13.3 建立客户签名的操作 515

13.3.1 驱动表 515

13.3.2 查找数据 518

13.3.3 最初的交易 520

13.3.4 旋转 521

13.3.5 总结 528

13.4 抽取特征 530

13.4.1 地理位置信息 530

13.4.2 日期时间列 531

13.4.3 字符串中的模式 532

13.5 总结客户行为 534

13.5.1 计算时间序列的斜率 534

13.5.2 周末消费者 537

13.5.3 下降的使用行为 540

13.6 小结 541

# 第14章  性能问题：高效使用SQL543

14.1 查询引擎和性能 544

14.1.1 用于理解性能的时间

复杂度 544

14.1.2 一个简单的示例 545

14.1.3 与性能相关的思考 547

14.1.4 性能的含义和测量 549

14.1.5 性能提升入门 549

14.2 高效使用索引 553

14.2.1 什么是索引？ 553

14.2.2 索引的简单示例 557

14.2.3 索引的限制 560

14.2.4 高效使用复合索引 562

14.3 何时使用OR是低效的？ 566

14.3.1 有时UNION ALL比OR

更好 566

14.3.2 有时LEFT OUTER JOIN比

OR更高效 567

14.3.3 有时多个条件表达式

更好 568

14.4 赞成和反对：表达一件事情的

不同方法 569

14.4.1 在Orders表中，哪些州

没有被识别？ 569

14.4.2 一个关于GROUP BY的

难题 571

14.4.3 小心COUNT(\*)=0 573

14.5 窗口函数 576

14.5.1 窗口函数适用于什么

地方？ 576

14.5.2 窗口函数的灵活使用 576

14.6 小结 582

附录  数据库之间的等价结构583

[显示部分信息](http://product.dangdang.com/javascript:void(0);)

前　　言

前 言本书的第1版使用我们熟悉的工具SQL和Excel，从实用的角度解释数据分析。这本书的指导原则是从问题出发，同时从业务角度和技术角度提供解决方案，以指导读者。这个方法被证明是非常成功的。从第1版到现在已经过去了10年，这期间已经发生了很多变化，工具本身也发生了很多变化。例如，当年的Excel还没有功能区，而且在当时的数据库中，窗口函数也非常罕见。一些工具，如Python和R，以及NoSQL数据库变得越来越常见，它们改变了分析师赖以生存的工具世界。然而，随着技术延伸到大大小小的各项业务中，关系型数据库在今天仍然被广泛使用，而且SQL也变得更加至关重要。对于很多商务人士，Excel工具仍然是做报表和展示的理想之选。大数据不再是未知的领域，它是我们每天都会面临的问题、挑战和机遇。根据底层软件的变化，在第2版中对本书的内容做了调整和更新，同时包含了更多的示例和技术，以及增加了关于数据库性能的一整章新内容。同时，我一直在努力保持本书第1版的优势。本书仍然围绕着数据、分析和展示的原则—— 少见地将三个功能放在一起处理。示例围绕着所提出的问题，同时讨论了这些问题的业务相关性和技术实现。示例使用的是真实的代码。数据、代码以及Excel示例都可以在配套网站上找到。撰写这本书的最初动机来源于我的一个同事—— Nick Drake，他是受过培训的统计学家。曾经，他一直在寻找一本书，关于介绍如何使用SQL编写可用于数据分析的复杂查询。当时，基于SQL的书籍，要么介绍SQL的基础查询结构，要么介绍数据库的工作原理。严格地讲，没有从分析数据的角度介绍SQL的书籍，也没有基于回答数据问题的书籍。在统计学的众多书籍中，没有一本书能够面对这样一个事实提出解决方案：统计学所用的数据，多数都存储于关系型数据库中，而本书则填补了这一空白。笔者与Michael Berry一起撰写的其他关于数据挖掘的书籍，侧重于高级算法和案例学习。相比之下，本书侧重于“操作方式”。首先描述了存储在数据库中的数据，然后继续完成准备数据和生成结果集的过程。书中穿插的内容，是我在这个领域多年经验的结晶，解释了结果集被应用的可能方式，以及为什么有些事情有效果，而有些事情无效。书中示例非常具有实践性，它们所使用的数据都在本书的配套网站上(www.wiley.com/go/dataanaly- sisusingsqlandexcel2e)。关于数据仓库和分析数据库的一个老生常谈的话题是它们实际上没有做任何事。是的，它们存储数据，能够将不同来源的数据汇集在一起，并整理数据使数据变得清晰。是的，它们定义业务维度，存储关于客户的事务，还可能总结重要的数据(是的，所有这些都非常重要！)然而，数据库中的数据存储在旋转的硬盘上，而且数据在计算机内存中的数据结构非常复杂。对于如此多的数据，信息却很少。我们如何探索这些数据(特别是描述客户的数据)？很多关于统计学建模和数据挖掘的华丽算法都有一条简单的规则：“无用输入，无用输出”。即使是最复杂的技术，也只有当数据是好数据时，结果才是好的。数据是理

# 其他

 SQL查询 16

1.4.1 做什么，而不是怎么去做 16

1.4.2 SELECT语句 17

1.4.3 一个基础的SQL查询 17

1.4.4 一个基本的SQL求和查询 19

1.4.5 联接表的意义 20

1.4.6  SQL的其他重要功能 26

1.5  子查询和公用表表达式 29

1.5.1 用于命名变量的子查询 29

1.5.2 处理统计信息的子查询 32

1.5.3 子查询和IN 33

1.5.4 用于UNION ALL的子查询 37

1.6  小结 38

# 第1讲 大数据简介 3

1.1　什么是大数据 4  
  
　　1.1.1 大数据的优势 5  
  
　　1.1.2 挖掘各种大数据源 6  
  
1.2　数据管理的历史——大数据的演化 7  
  
1.3 大数据的结构化 9  
  
1.4 大数据要素 13  
  
　　1.4.1 数据量 13  
  
　　1.4.2 速度 14  
  
　　1.4.3 多样性 14

# 线性回归分析 118

1. 4.1　简单线性回归模型 119  
   4.1.1　回归模型和回归方程 119  
   4.1.2　估计的回归方程 120  
   4.2　最小二乘法 121  
   4.3　简单线性回归模型的拟合效果 126  
   4.3.1　离差平方和的分解 126  
   4.3.2　可决系数 128  
   4.3.3　Excel可决系数计算 129  
   4.4　多元回归模型 130  
   4.4.1　多元回归模型和多元回归方程 130  
   4.4.2　估计的多元回归方程 130  
   4.4.3　最小二乘法和多元回归 130  
   4.4.4　多元回归分析实例 131  
   4.4.5　Excel中的多元回归求解 132  
   4.5　回归推断分析 135  
   4.5.1　推断分析的必要条件 135  
   4.5.2　总体回归关系检验 139  
   4.5.3　回归参数检验 140  
   4.5.4　不显著自变量处理 142  
   4.5.5　多重共线性 143  
   4.5.6　大样本情形 145  
   4.6　属性自变量 149  
   4.6.1　引入属性自变量 149  
   4.6.2　引入属性变量后回归参数的意义 151  
   4.6.3　多个属性变量的处理 152  
   4.7　非线性回归模型 153

# 时间序列分析与预测 182

1. 5.1　时间序列的几种类型 184  
   5.1.1　水平变化状态的时间序列 185  
   5.1.2　带有趋势的时间序列 186  
   5.1.3　带有季节性波动的时间序列 188  
   5.1.4　同时带有趋势和季节性波动的时间序列 189  
   5.1.5　带有周期性波动的时间序列 190  
   5.1.6　如何识别时间序列形态 190  
   5.2　预测精度问题 190  
   5.3　移动平均与指数平滑法 194  
   5.3.1　移动平均法 195  
   5.3.2　指数平滑法 198  
   5.4　回归预测分析 202  
   5.4.1　线性趋势回归分析 202  
   5.4.2　带有季节性效应的回归分析 204  
   5.4.3　因果关系的回归分析预测 208  
   5.4.4　存在因果变量和趋势及季节效应的回归预测 211  
   5.5　预测模型优良性评估 211

# 第6章　数据挖掘 226

6.1　数据抽样 227  
6.2　数据预处理 228  
6.2.1　缺失数据问题 228  
6.2.2　识别异常值和错误数据 229  
6.2.3　代表性变量 229  
6.3　无指导学习 230  
6.3.1　聚类分析 230  
6.3.2　关联规则 239  
6.4　指导学习 242  
6.4.1　数据分割 243  
6.4.2　分类准确度 246  
6.4.3　预测准确度 249  
6.4.4　k最近邻算法 250  
6.4.5　分类回归树 254  
6.4.6　逻辑回归 268

# 第12章　决策分析 477

12.1　问题的表述 478  
12.1.1　报偿表 479  
12.1.2　决策树 479  
12.2　不使用概率的决策分析 480  
12.2.1　乐观主义准则 481  
12.2.2　保守主义准则 481  
12.2.3　后悔主义准则 482  
12.3　使用概率的决策分析 483  
12.3.1　期望值准则 483  
12.3.2　风险分析 485  
12.3.3　敏感性分析 486  
12.4　运用样本信息的决策分析 487  
12.4.1　追加样本信息的决策分析 487  
12.4.2　样本信息的期望值 491  
12.4.3　完全信息期望值 492  
12.5　利用贝叶斯定理计算状态枝概率 493  
12.6　效用决策 495  
12.6.1　引言 495  
12.6.2　效用与决策分析 496  
12.6.3　效用函数 500  
12.6.4　指数效用函数 502  
本章小结 503

# 参考资料

## 《深入浅出数据分析》(（美）米尔顿　著)【简介\_书评\_在线阅读】 - 当当图书.mhtml

## 《数据分析技术(第2版) 使用SQL和Excel工具》([美]Gordon S. Linoff 著 陶佰明 译)【简介\_书评\_在线阅读】 - 当当图书.mhtml

本书的14章可以分为4部分。前3章介绍SQL、Excel和统计学的核心概念。中间7章讨论特别适合使用SQL和Excel的数据探索和数据分析技术。在后续的3章中，从统计学和数据挖掘的角度，介绍了关于建模的更正式的思想。\*后，新增的第14章讨论编写SQL查询时的性能问题。每一章都通过不同的视角，介绍使用SQL和Excel做数据分析的方方面面，包括：● 使用数据分析的基础示例● 分析师需要回答的问题● 详解数据分析技术的工作原理● 实现技术的SQL语法● 以表格或图表展示结果，以及如何在Excel中创建它们

第1章 数据挖掘者眼中的SQL 1

第2章 表中有什么？开始数据探索39

第3章 不同之处是如何不同？ 79

第4章 发生的地点在何处？ 119

第5章 关于时间 161

第6章 客户的持续时间有多久？使用

第7章 影响生存率的因素：客户

第8章 多次购买以及其他重复事件303

第9章 购物车里有什么？购物车

第10章 关联规则 387

第11章 SQL数据挖掘模型 423

第12章 最佳拟合线：线性回归

第13章 为进一步分析数据创建客户

第14章 性能问题：高效使用SQL543

前 言本书的第1版使用我们熟悉的工具SQL和Excel，从实用的角度解释数据分析。这本书的指导原则是从问题出发，同时从业务角度和技术角度提供解决方案，以指导读者。这个方法被证明是非常成功的。从第1版到现在已经过去了10年，这期间已经发生了很多变化，工具本身也发生了很多变化。例如，当年的Excel还没有功能区，而且在当时的数据库中，窗口函数也非常罕见。一些工具，如Python和R，以及NoSQL数据库变得越来越常见，它们改变了分析师赖以生存的工具世界。然而，随着技术延伸到大大小小的各项业务中，关系型数据库在今天仍然被广泛使用，而且SQL也变得更加至关重要。对于很多商务人士，Excel工具仍然是做报表和展示的理想之选。大数据不再是未知的领域，它是我们每天都会面临的问题、挑战和机遇。根据底层软件的变化，在第2版中对本书的内容做了

## 《大数据分析师权威教程 大数据分析与预测建模》(Wrox国际IT认证项目组)【简介\_书评\_在线阅读】 - 当当图书.mhtml

ine 454: 模块1　大数据入门

Line 600: 模块2　分析和R编程入门

Line 806: 模块3 使用R进行数据分析

Line 1032: 模块4 使用R进行高级分析

第1讲 大数据简介 3

第2讲 大数据在商业上的应用 23

第3讲 处理大数据的技术 43

第4讲 了解Hadoop生态系统 59

第5讲 MapReduce基础 77

第1讲 理解分析 97

第2讲 分析方法与工具 116

第3讲 探索R 136

第4讲 将数据集读入R，从R导出数据 154

第5讲 在R中操纵和处理数据 177

第1讲 使用R中的函数和包 207

第2讲 R中的描述性统计 231

第3讲 用函数、循环和数据帧分析数据 270

第4讲 R中的图形分析 286

第5讲 R中的假设检验 325

第1讲 R中的线性回归 357

第2讲 非线性回归 390

第3讲 聚类分析 419

第4讲 决策树 440

第5讲 R和Hadoop的集成及Hive介绍 465

## 《商业数据分析》([美]杰弗里 D.坎姆（Jeffrey D.Camm）等)【简介\_书评\_在线阅读】 - 当当图书.mhtml

Line 374: 第1章　导论 1

Line 391: 本章小结 12

Line 393: 第2章　描述统计分析 15

Line 430: 本章小结 52

Line 435: 第3章　数据可视化 66

Line 456: 本章小结 101

Line 461: 第4章　线性回归分析 118

Line 495: 本章小结 165

Line 500: 第5章　时间序列分析与预测 182

Line 518: 本章小结 212

Line 523: 第6章　数据挖掘 226

Line 539: 本章小结 276

Line 543: 第7章　电子表格模型 285

Line 561: 本章小结 305

Line 565: 第8章　线性优化模型 314

Line 584: 本章小结 344

Line 590: 第9章　整数线性优化 356

Line 604: 本章小结 377

Line 609: 第10章　非线性优化问题 391

Line 622: 本章小结 408

Line 627: 第11章　Monte Carlo模拟 420

Line 641: 本章小结 454

Line 647: 第12章　决策分析 477

Line 669: 本章小结 503

Line 679: 本书一共由12章组成，其中：