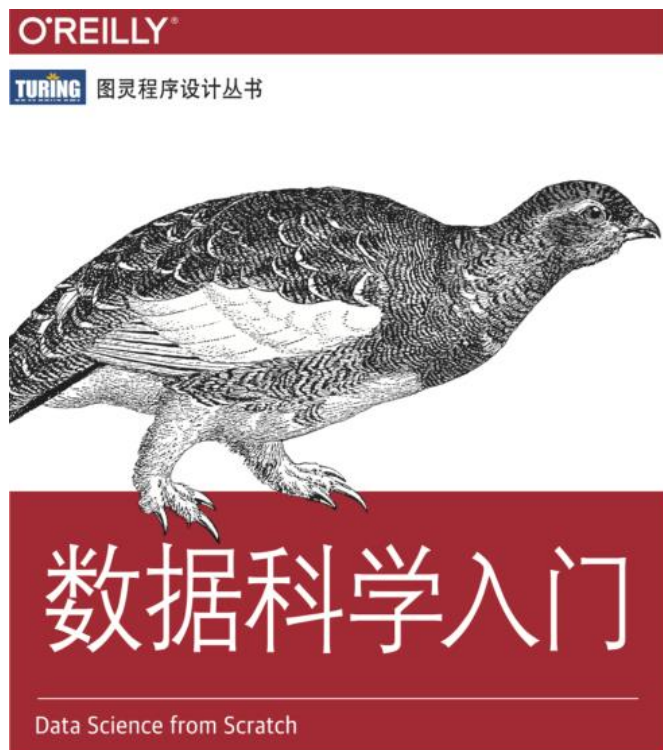


# 应用数据科学导论

**Ht\_song@163.com**

**2020.02**

# 参考书



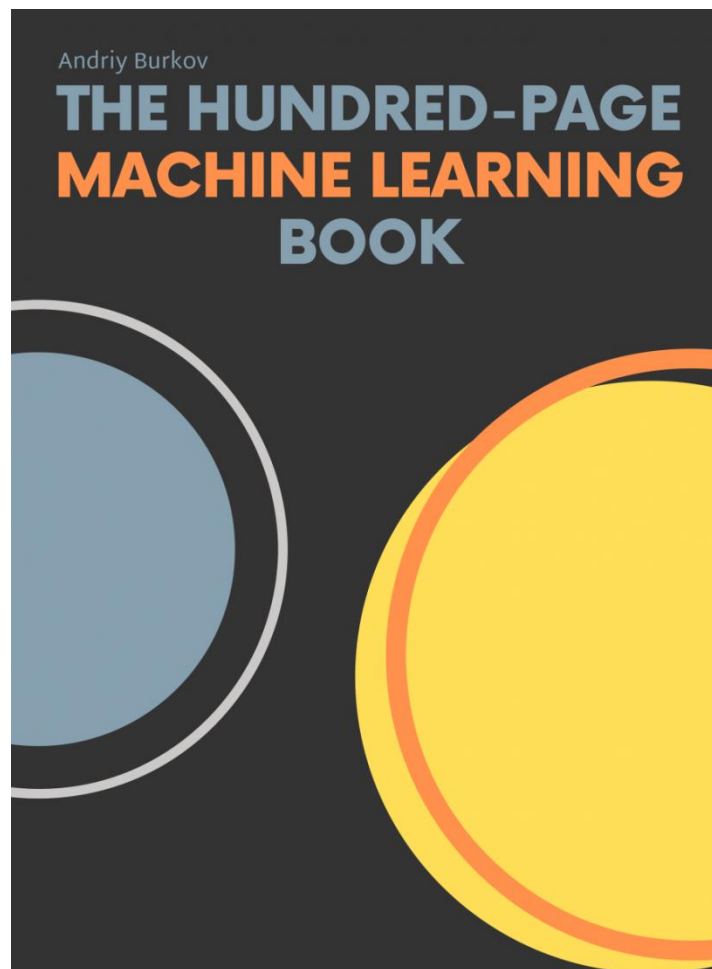
[美] Joel Grus 著  
高蓉 韩波 译



中国工信出版集团



人民邮电出版社  
POSTS & TELECOM PRESS



# 主要内容

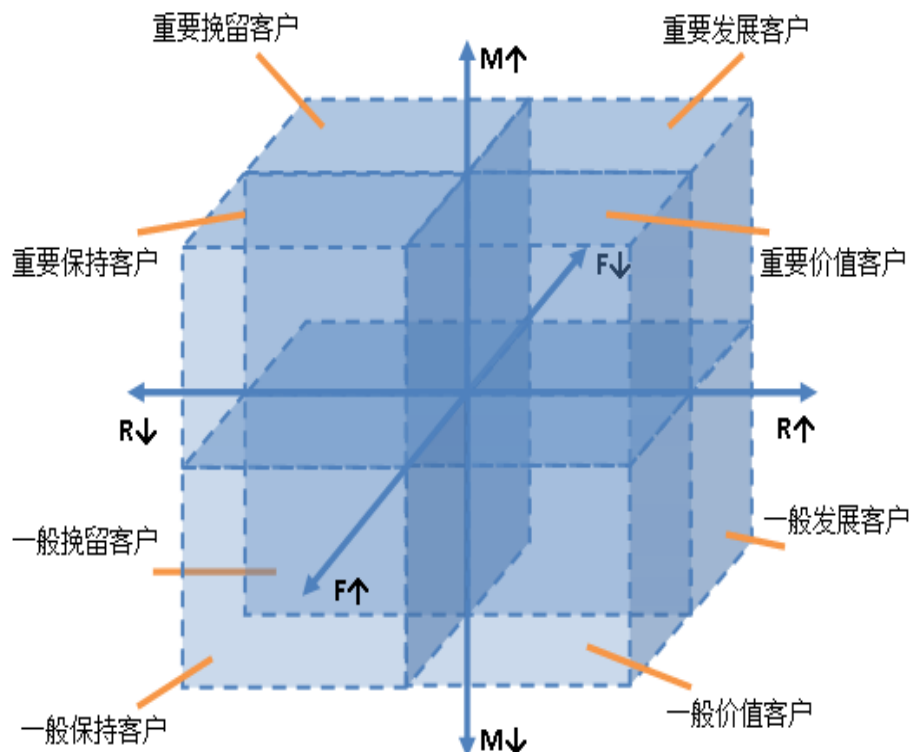
- 数据科学简介
- Python语言
- 数据统计可视化
- 数据分析方法
- 分析项目实践——综合应用

# 传统方法存在的缺陷

- 广泛用于分析客户价值的是RFM模型，它是通过三个指标（最近消费时间间隔(Recency)、消费频率(Frequency)、消费金额(Monetary)）来进行客户细分，识别出高价值的客户。如果分析航空公司客户价值，此模型不再适用，存在一些缺陷和不足：

一：在模型中，消费金额表示在一段时间内，客业产品金额的总和。因航空票价受到运输距离、舱位等级等多种因素影响，同样消费金额的不同旅客对航空公司的价值是不同的。因此这个指标并不适合用于航空公司的客户价值分析。

二：传统模型分析是利用属性分箱方法进行分析如图，但是此方法细分的客户群太多，需要一一识别客户特征和行为，提高了针对性营销的成本。



# 原始数据情况

- 客户信息属性说明
- 航空客户信息数据

# 原始数据情况

- 客户信息属性说明，针对航空客户的信息，对每个属性进行相应说明。

	属性名称	属性说明
客户基本信息	MEMBER_NO	会员卡号
	FFP_DATE	入会时间
	FIRST_FLIGHT_DATE	第一次飞行日期
	GENDER	性别
	FFP_TIER	会员卡级别
	WORK_CITY	工作地城市
	WORK_PROVINCE	工作地所在省份
	WORK_COUNTRY	工作地所在国家
乘机信息	AGE	年龄
	FLIGHT_COUNT	观测窗口内的飞行次数
	LOAD_TIME	观测窗口的结束时间
	LAST_TO_END	最后一次乘机时间至观测窗口结束时长
	AVG_DISCOUNT	平均折扣率
	SUM_YR	观测窗口的票价收入
	SEG_KM_SUM	观测窗口的总飞行公里数
	LAST_FLIGHT_DATE	末次飞行日期
积分信息	AVG_INTERVAL	平均乘机时间间隔
	MAX_INTERVAL	最大乘机间隔
	EXCHANGE_COUNT	积分兑换次数
	EP_SUM	总精英积分
	PROMOPTIVE_SUM	促销积分
	PARTNER_SUM	合作伙伴积分
	POINTS_SUM	总累计积分
	POINT_NOTFLIGHT	非乘机的积分变动次数
	BP_SUM	总基本积分

以过去某个时间点为结束时间，某一时间长度作为宽度，得到历史时间范围内的一个时间段。

# 原始数据情况

- 航空客户信息，其中已经包含会员档案信息和其乘坐航班记录等

MEMBER	FFP_DATE	FIRST_FLIG	GENDE	FFP_TI	WORK_CIT	WORK_PROVIN	WORK_COU	AGE	LOAD_TIME	FLIGH	BP_SUM
54993	2006/11/02	2008/12/24	男	6	.	北京	CN	31	2014/03/31	210	505308
28065	2007/02/19	2007/08/03	男	6		北京	CN	42	2014/03/31	140	362480
55106	2007/02/01	2007/08/30	男	6	.	北京	CN	40	2014/03/31	135	351159
21189	2008/08/22	2008/08/23	男	5	Los Angele	CA	US	64	2014/03/31	23	337314
39546	2009/04/10	2009/04/15	男	6	贵阳	贵州	CN	48	2014/03/31	152	273844
56972	2008/02/10	2009/09/29	男	6	广州	广东	CN	64	2014/03/31	92	313338
44924	2006/03/22	2006/03/29	男	6	乌鲁木齐市	新疆	CN	46	2014/03/31	101	248864
22631	2010/04/09	2010/04/09	女	6	温州市	浙江	CN	50	2014/03/31	73	301864
32197	2011/06/07	2011/07/01	男	5	DRANCY		FR	50	2014/03/31	56	262958
31645	2010/07/05	2010/07/05	女	6	温州	浙江	CN	43	2014/03/31	64	204855
58877	2010/11/18	2010/11/20	女	6	PARIS	PARIS	FR	34	2014/03/31	43	298321
37994	2004/11/13	2004/12/02	男	6	北京	.	CN	47	2014/03/31	145	256093
28012	2006/11/23	2007/11/18	男	5	SAN MARI	CA	US	58	2014/03/31	29	210269
54943	2006/10/25	2007/10/27	男	6	深圳	广东	CN	47	2014/03/31	118	241614
57881	2010/02/01	2010/02/01	女	6	广州	广东	CN	45	2014/03/31	50	289917
1254	2008/03/28	2008/04/05	男	4	BOWLAND	CALIFORNIA	US	63	2014/03/31	22	286164
8253	2010/07/15	2010/08/20	男	6	乌鲁木齐	新疆	CN	48	2014/03/31	101	219995
58899	2010/11/10	2011/02/23	女	6	PARIS		FR	50	2014/03/31	40	249882
26955	2006/04/06	2007/02/22	男	6	乌鲁木齐市	新疆	CN	54	2014/03/31	64	215013
41616	2011/08/29	2011/10/22	男	6	东莞	广东	CN	41	2014/03/31	38	191038
21501	2008/07/30	2008/11/21	男	6	.	北京	CN	49	2014/03/31	106	220641
41281	2011/06/07	2011/06/09	男	6	VECHEL	NORD BRABAN	AN		2014/03/31	23	255573
47229	2005/04/10	2005/04/10	男	6	广州	广东	CN	69	2014/03/31	94	193169
28474	2010/04/13	2010/04/13	男	6		CA	US	41	2014/03/31	20	256337
58472	2010/02/14	2010/03/01	女	5			FR	48	2014/03/31	44	204801
13942	2010/10/14	2010/11/01	男	6	PARIS	FRANCE	FR	39	2014/03/31	62	241719
45075	2007/02/01	2007/03/23	男	6	湛江	广东	CN	46	2014/03/31	213	217809

# 挖掘目标

1. 借助航空公司客户数据，对客户进行分类；
2. 对不同的客户类别进行特征分析，比较不同类客户的客户价值；
3. 对不同价值的客户类别提供个性化服务，制定相应的营销策略。



# 目录

1	背景与挖掘目标
2	分析方法与过程
3	上机实验
4	拓展思考

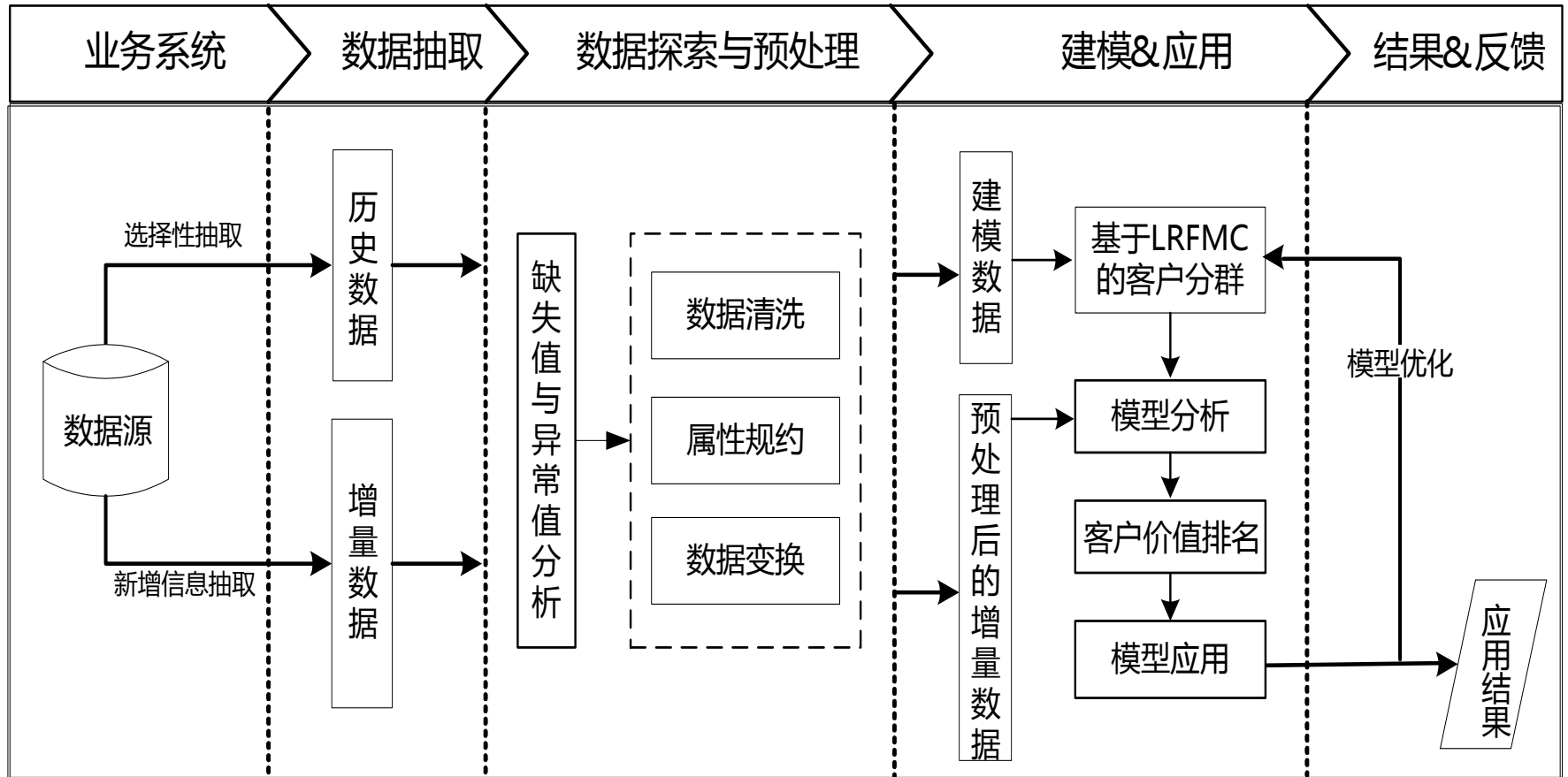
# 分析方法与过程

初步分析：提出适用航空公司的LRFMC模型

- 因消费金额指标在航空公司中不适用，故选择客户在一定时间内累积的飞行里程 $M$ 和客户乘坐舱位折扣系数的平均值 $C$ 两个指标代替消费金额。此外，考虑航空公司会员加入时间在一定程度上能够影响客户价值，所以在模型中增加客户关系长度 $L$ ，作为区分客户的另一指标，因此构建出LRFMC模型。
- 采用聚类的方法对客户进行细分，并分析每个客户群的特征，识别其客户价值。

# 分析方法与过程

总体流程：



# 分析方法与过程

## 第1步：数据抽取

- 以2014-03-31为结束时间，选取宽度为两年的时间段作为分析观测窗口，抽取观测窗口内有乘机记录的所有客户的详细数据形成历史数据。对于后续新增的客户详细信息，利用其数据中最大的某个时间点作为结束时间，采用上述同样的方法进行抽取，形成增量数据。
- 根据末次飞行日期，从航空公司系统内抽取2012-04-01至2014-03-31内所有乘客的详细数据，总共62988条记录。

# 分析方法与过程

## 第2步：探索分析

- 原始数据中存在票价为空值，票价为空值的数据可能是客户不存在乘机记录造成。
- 票价最小值为0、折扣率最小值为0、总飞行公里数大于0的数据。其可能是客户乘坐0折机票或者积分兑换造成。

属性名称	SUM_YR_1	SUM_YR_2	...	SEG_KM_SUM	AVG_DISCOUNT
空值记录数	551	138	...	0	0
最大值	239560	234188	...	580717	1.5
最小值	0	0	...	368	0



编程练习

# 分析方法与过程

## 第3步：数据预处理

1. 数据清洗：从业务以及建模的相关需要方面考虑，筛选出需要的数据
  - a) 丢弃票价为空的数据。
  - b) 丢弃票价为0、平均折扣率不为0、总飞行公里数大于0的数据。



编程练习

# 分析方法与过程

## 第3步：数据预处理

- 属性规约：原始数据中属性太多，根据LRFMC模型，选择与其相关的六个属性，删除不相关、弱相关或冗余的属性。

LOAD_TIME	FFP_DATE	LAST_TO_END	FLIGHT_COUNT	SEG_KM_SUM	AVG_DISCOUNT
2014/4/1	2006/03/31	6.6	3	18770	0.66
2014/4/1	2006/03/31	3.8	24	35087	0.62
2014/4/1	2006/04/07	2.8	9	20660	0.52
2014/4/1	2006/08/10	1	12	23071	0.51
2014/4/1	2008/02/07	3.17	3	2897	0.95
2014/4/1	2010/09/16	1.57	3	4608	0.65
2014/4/1	2011/04/28	17.83	2	3390	0.48
2014/4/1	2012/03/09	4.13	8	11797	1.35
2014/4/1	2012/08/31	5.9	6	6355	0.75
2014/4/1	2005/09/29	0.4	54	62170	0.79
2014/4/1	2009/01/23	2.33	24	15894	0.6
2014/4/1	2009/01/30	0.07	13	19517	0.72
2014/4/1	2009/01/30	4.63	10	12686	0.55
2014/4/1	2009/02/13	14.93	13	10992	1.33
2014/4/1	2009/04/25	10.27	3	4137	0.67
2014/4/1	2009/06/12	0.33	19	37415	0.63
2014/4/1	2010/03/25	1.63	13	24156	0.79
2014/4/1	2010/04/15	22.23	3	1559	0.87
2014/4/1	2010/05/20	23.17	2	1870	0.6
2014/4/1	2010/07/08	2.6	30	46621	0.93
2014/4/1	2011/03/10	4.57	4	7999	0.58

# 分析方法与过程

## 第3步：数据预处理

### 3. 数据变换

a) 属性构造

b) 数据标准化



# 分析方法与过程

## 第3步：数据预处理

### 3. 数据变换

a) **属性构造**：因原始数据中并没有直接给出LRFMC五个指标，需要构造这五个指标。

$L = \text{LOAD\_TIME} - \text{FFP\_DATE}$

会员入会时间距观测窗口结束的月数 = 观测窗口的结束时间 - 入会时间[单位：月]

$R = \text{LAST\_TO\_END}$

客户最近一次乘坐公司飞机距观测窗口结束的月数 = 最后一次乘机时间至观察窗口末端时长  
[单位：月]

$F = \text{FLIGHT\_COUNT}$

客户在观测窗口内乘坐公司飞机的次数 = 观测窗口的飞行次数[单位：次]

$M = \text{SEG\_KM\_SUM}$

客户在观测时间内在公司累计的飞行里程 = 观测窗口总飞行公里数[单位：公里]

$C = \text{AVG\_DISCOUNT}$

客户在观测时间内乘坐舱位所对应的折扣系数的平均值 = 平均折扣率[单位：无]

# 分析方法与过程

## 第3步：数据预处理

### 3. 数据变换

a) **数据标准化**：因五个指标的取值范围数据差异较大，为了消除数量级数据带来的影响，需要对数据进行标准化处理。

属性名称	L	R	F	M	C
最小值	12.23	0.03	2	368	0.14
最大值	114.63	24.37	213	580717	1.5



编程练习

# 分析方法与过程

## 第5步：构建模型

### 1. 构建航空客户价值分析模型

- a) 客户K-Means聚类
- b) 客户价值分析
- c) 模型应用

# 分析方法与过程

## 第5步：构建模型

### 1. 客户K-Means聚类

采用K-Means聚类算法对客户数据进行分群，将其聚成五类（需要结合业务的理解与分析来确定客户的类别数量）。

表 7-9 客户聚类结果

聚类 类别	聚类 个数	聚类中心				
		ZL	ZR	ZF	ZM	ZC
客户群 1	5337	0.483	-0.799	2.483	2.424	0.308
客户群 2	15735	1.160	-0.377	-0.087	-0.095	-0.158
客户群 3	12130	-0.314	1.686	-0.574	-0.537	-0.171
客户群 4	24644	-0.701	-0.415	-0.161	-0.165	-0.255
客户群 5	4198	0.057	-0.006	-0.227	-0.230	2.191

\*由于 K-Means 聚类是随机选择类标号，因此上机实验得到结果中的类标号可能与此不同



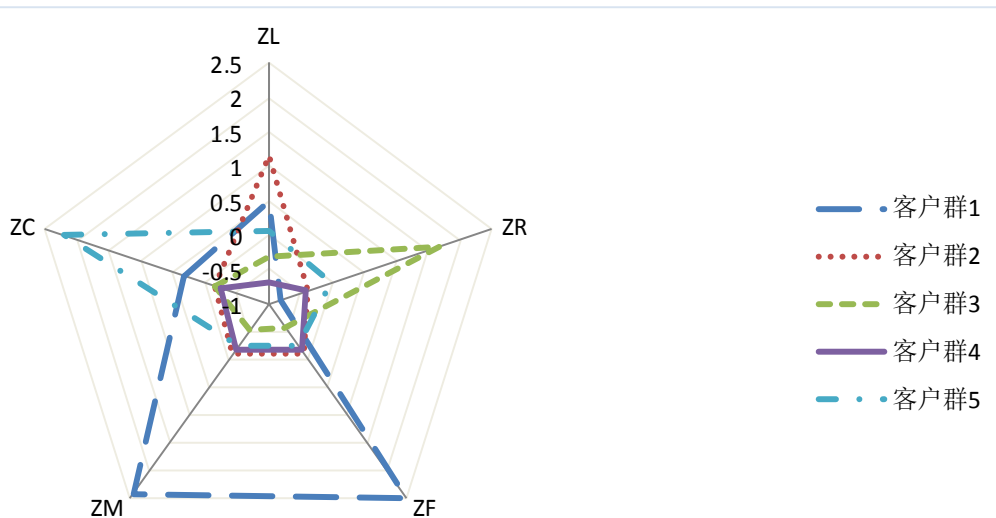
编程练习

# 分析方法与过程

## 第5步：构建模型

### 2. 客户价值分析

对聚类结果进行特征分析，其中客户群1在F、M属性最大，在R属性最小；客户群2在L属性上最大；客户群3在R属性上最大，在F、M属性最小；客户群4在L、C属性上最小；客户群5在C属性上最大。























# 分析方法与过程

## 第5步：构建模型

### 2. 客户价值分析

根据业务定义五个等级的客户类别：重要保持客户、重要发展客户、重要挽留客户、一般客户与低价值客户。

	重要保持客户	重要发展客户	重要挽留客户	一般客户与低价值客户
平均折扣系数 (C)				
最近乘机距今 的时间长度 (R)				
飞行次数 (F)				
总飞行里程 (M)				
会员入会时间 (L)				

# 分析方法与过程

## 第5步：构建模型

### 2. 客户价值分析

客户群价值排名：根据每种客户类型的特征，对各类客户群行客户价值排名，获取高价值客户信息。

客户群	排名	排名含义
客户群 1	1	重要保持客户
客户群 5	2	重要发展客户
客户群 2	3	重要挽留用户
客户群 4	4	一般客户
客户群 3	5	低价值客户

# 分析方法与过程

## 第5步：构建模型

3. 模型应用：根据各个客户群的特征，可采取一些营销手段和策略。

a) 会员的升级与保级。

b) 首次兑换。

c) 交叉销售。



# 目录

1	背景与挖掘目标
2	分析方法与过程
3	上机实验
4	拓展思考

# 上机实验

## 1. 实验目的

- a) 了解K-Means聚类算法在客户价值分析实例中的应用。
- b) 利用Python实现数据z-score (标准差) 标准化以及模型的K-Means聚类过程。

## 2. 实验内容

- a) 利用程序，读入LRFMC指标文件，分别计算各个指标的均值与其标准差，使用标准差标准化公式完成LRFMC指标的标准化，并将标准化后的数据进行保存。
- b) 编写MATLAB程序，完成客户的K-Means聚类，获得聚类中心与类标号。输出聚类中心的特征图，并统计每个类别的客户数。

# 上机实验拓展

## 1. 实验目的

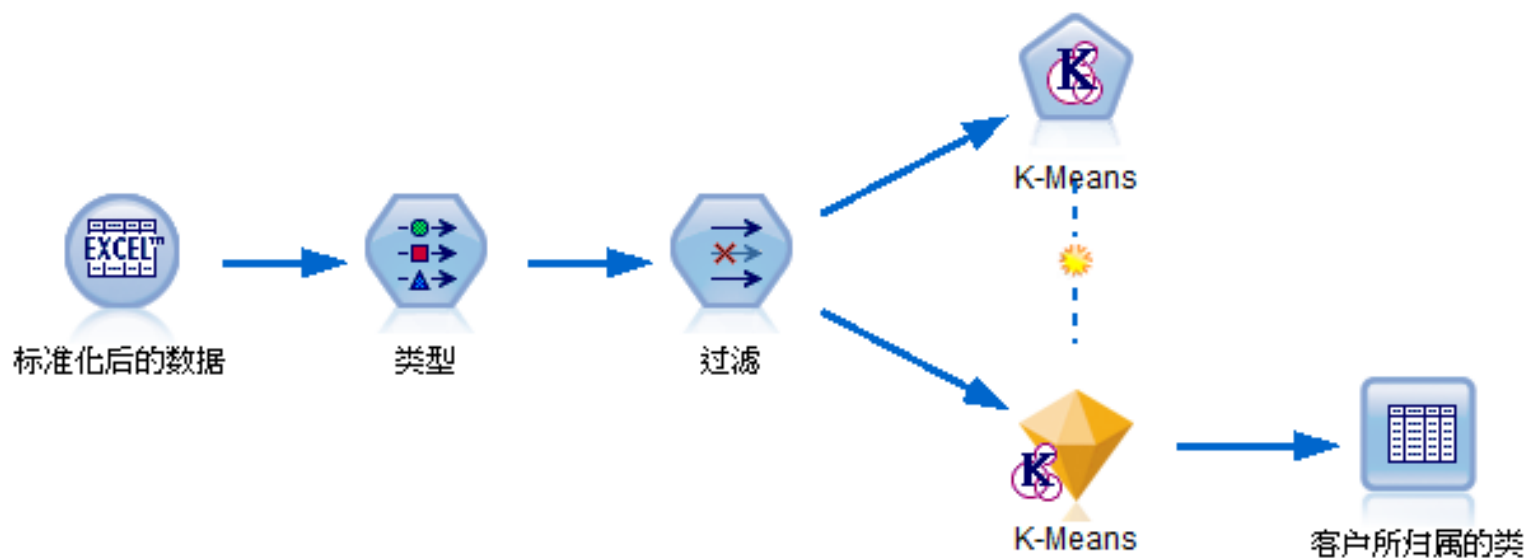
a) 掌握K-Means聚类算法构建模型

## 2. 实验内容

a) 对使用标准差标准化公式后的LRFMC指标进行聚类，设置聚类参数为5类，数据见“上机实验拓展/data/model.xls”，使用K-Means聚类算法实现航空公司客户细分模型，获得聚类中心与类标号，并统计每个类别的客户数。

# 上机实验拓展



## a) K-Means聚类



注：运行模型文件时需重新设置数据文件路径

# 上机实验拓展

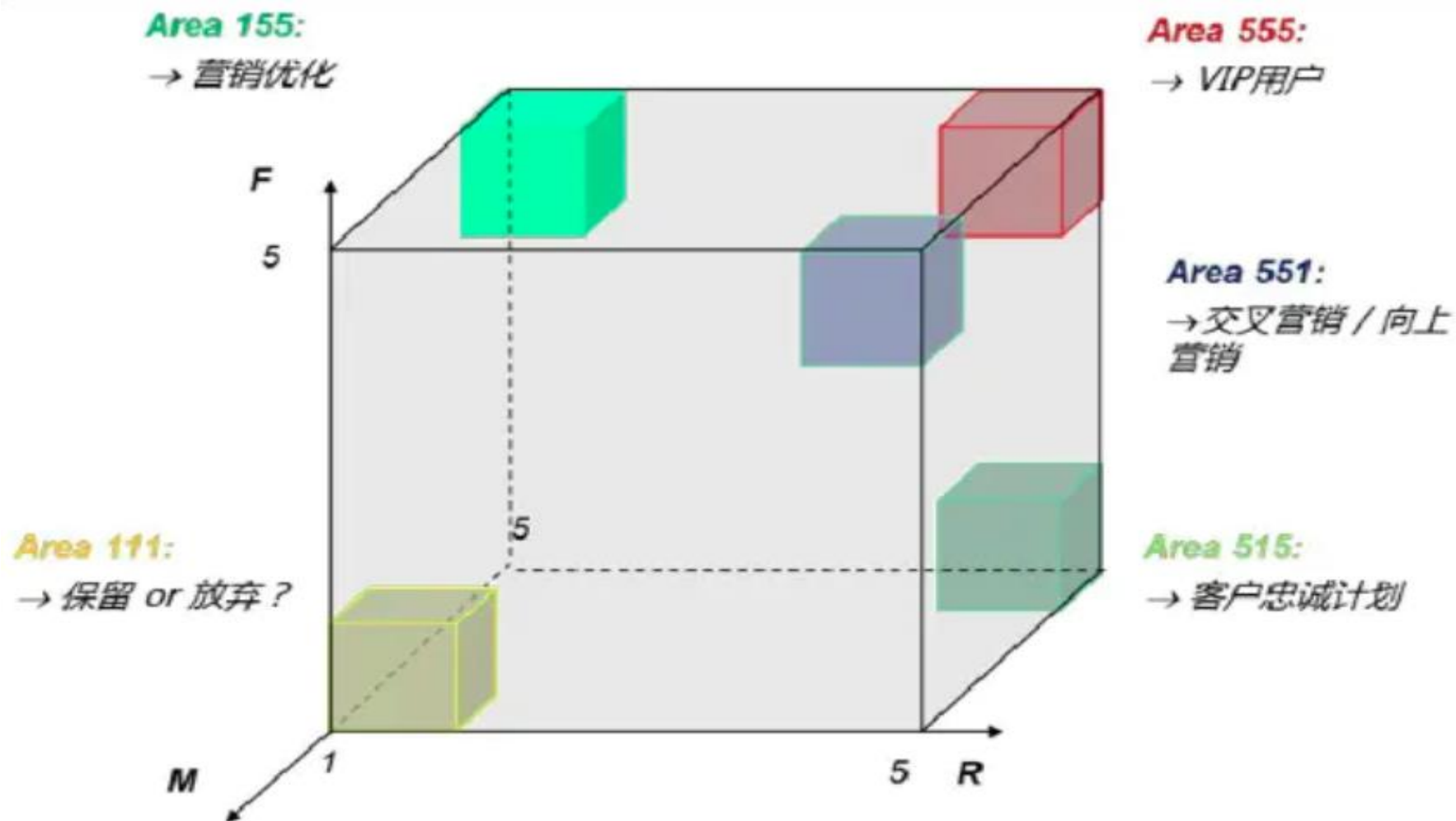
其聚类结果：

聚类	聚类-5	聚类-3	聚类-2	聚类-1	聚类-4
标签					
说明					
大小	 35.3% (21874)	 22.0% (13633)	 19.1% (11849)	 15.9% (9863)	 7.8% (4825)
输入	F -0.02	F 0.20	F 0.47	F -0.55	F -0.52
	L -0.86	L 0.28	L 1.48	L -0.75	L 1.00
	M -0.01	M 0.17	M 0.44	M -0.50	M -0.49
	R -0.44	R -0.49	R -0.55	R 1.60	R 1.49
	C -0.09	C 0.05	C 0.16	C -0.06	C 0.02

# 目录

1	背景与挖掘目标
2	分析方法与过程
3	上机实验
4	拓展思考

# 分类之后做什么？



# 拓展思考

- 在国内航空市场竞争日益激烈的背景下，客户流失问题是影响公司利益的重要因素之一。如何如何改善流失问题，继而提高客户满意度、忠诚度，维护自身的市场和利益？
- 客户流失分析可以针对目前老客户进行分类预测。针对航空公司客户信息数据附件（见：`/air_data.csv`）可以进行老客户以及客户类型的定义（例如：将其中将飞行次数大于6次的客户定义为老客户，已流失客户定义为：第二年飞行次数与第一年飞行次数比例小于50%的客户等）。
- 选取客户信息中的关键属性如：会员卡级别，客户类型（流失、准流失、未流失），平均折扣率，积分兑换次数，非乘机积分总和，单位里程票价，单位里程积分等。通过这些信息构建客户的流失模型，运用模型预测未来客户的类别归属（未流失、准流失，或已流失）。



# 数据，是信息时代的真相！