CDA数据分析师毕业答辩

第五组

成员：陈　涛　彭飞舞

肖　威　刘　娜

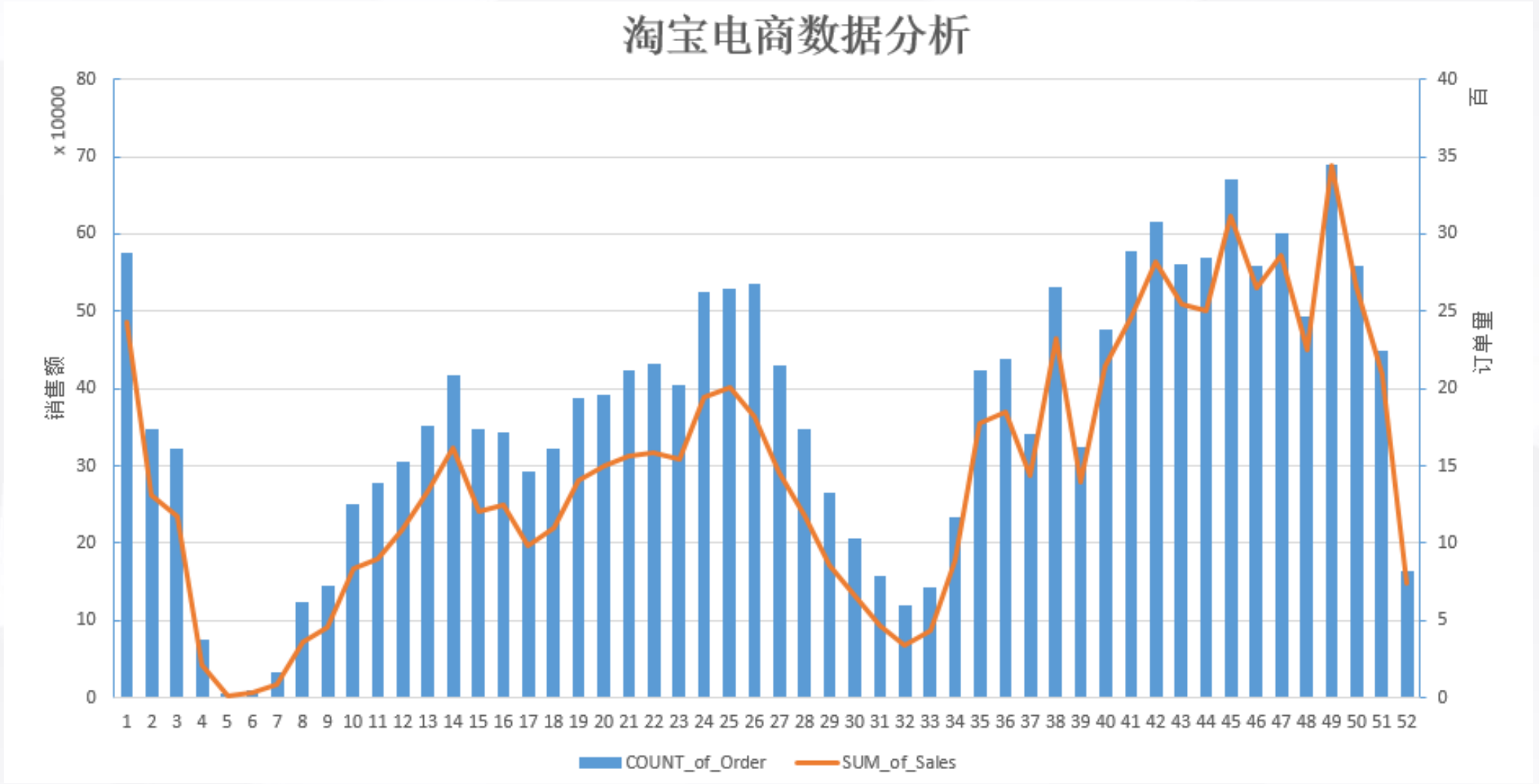
2015年11月25日星期三

数据分析的力量

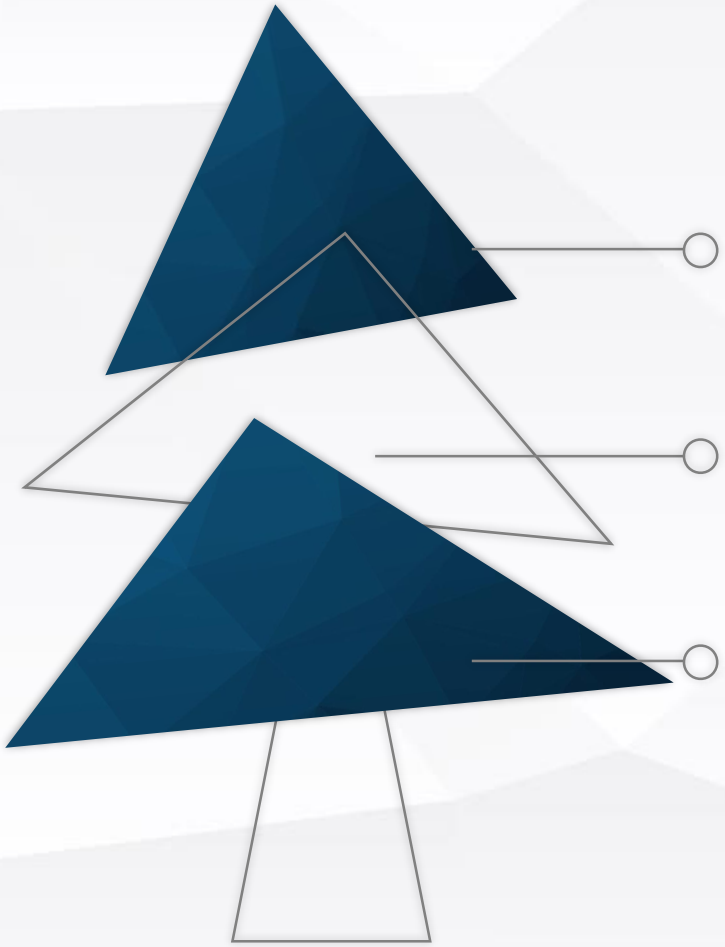
- 在购买不同的商品时，消费者如何分配他们的支出?
- 社会收入花费了多少，又节省了多少?
- 如何才能最好地测量与分析福利与贫穷？

普林斯顿大学教授安格斯·迪顿（Angus Deaton）
2015年诺贝尔经济学奖得者

淘宝电商数据分析与挖掘



目录



理解业务与数据

挑选商品 → 提交订单 → 支付订单 → 卖家发货

订单数据

- 买家信息
- 商品信息
- 支付信息
- 其他信息

CONTENTS PROCEDURE

按字母排序的变量和属性列表

#	变量	类型	长度	输出格式	输入格式	标签
13	Address	字符	40	$CHAR39.	$CHAR39.	收货地址
23	City	字符	40			城市
3	Count_ID	字符	9	$CHAR9.	$CHAR9.	买家支付宝账号
24	Create_order_date	数值	8	DATE9.		订单创建日期
12	Message_num	数值	8	BEST12.	BEST12.	买家留言字数
2	Name	字符	11	$CHAR11.	$CHAR11.	买家会员名
1	Order	数值	8	BEST16.	COMMA12.	订单编号
20	Order_note	数值	8	BEST12.	BEST12.	订单备注
11	Order_status	字符	8	$CHAR8.	$CHAR8.	订单状态
9	Pay_amount	数值	8	BEST12.	BEST12.	买家实际支付金额
25	Pay_date	数值	8	DATE9.		订单支付日期
10	Pay_integration	数值	8	BEST12.	BEST12.	买家实际支付积分
16	Pay_time	数值	8	DATETIME18.	DATETIME18.	订单付款时间
17	Product_name	字符	2680	$CHAR1782.	$CHAR1782.	宝贝标题
21	Product_num	数值	8	BEST12.	BEST12.	宝贝总数量
18	Product_type	数值	8	BEST12.	BEST12.	宝贝种类
22	Province	字符	40			省份
4	Sales	数值	8	BEST12.	BEST12.	买家应付货款
7	Total_amount	数值	8	BEST12.	BEST12.	总金额

	观测	120757
	变量	36
	索引	0
	观测长度	8416
	删除的观测	0
	已压缩	NO
	已排序	NO
(EUC)		

继续理解数据

甄别无用变量

- 支付积分
- 返点积分
- 实际支付积分
- 宝贝种类
- ……

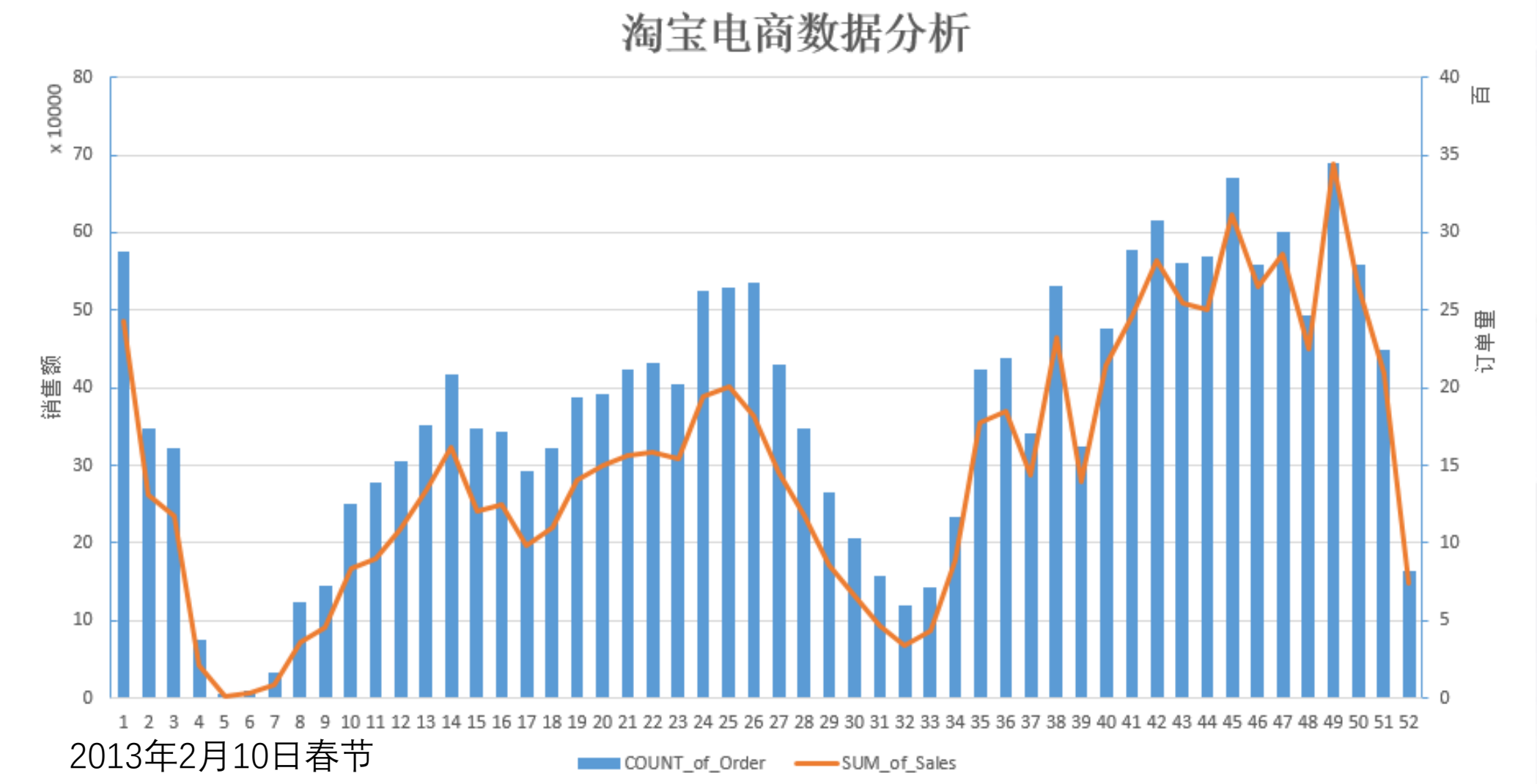
生成需要变量

- 从收货地址中提取省份
- 买家支付账户个数
- 提交、支付订单的时间间隔
- 购买频率
- ……

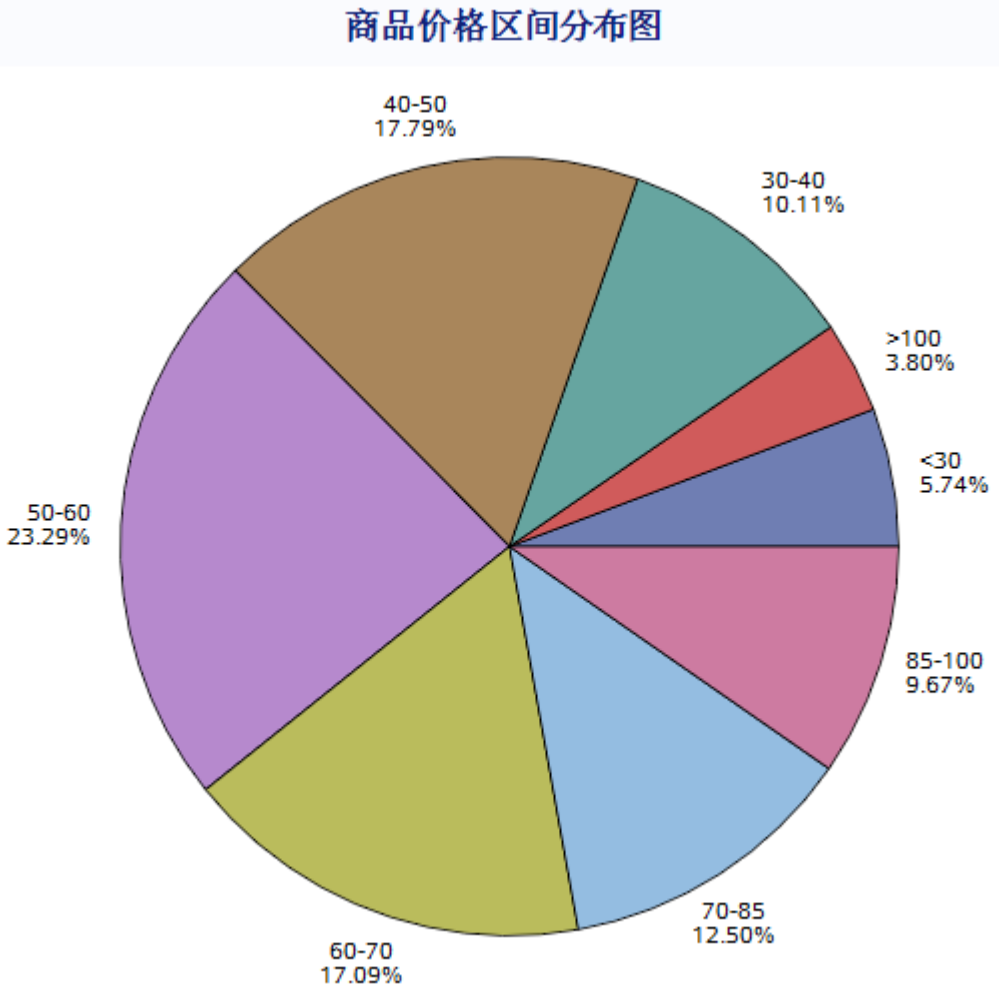
```
/*创建变量*/
data bysj.bysj;
  set bysj.bysj;
  Province = scan(address,1,' ');
  City = scan(address,2,' ');
  Create_order_date = datepart(create_order_time);
  Pay_date = datepart(Pay_time);
  timed = pay_date - create_order_date;
  timeh = (pay_time - create_order_time)/60;
  create_order_hour = hour(create_order_time);
  month = month(create_order_date);
  week = int((Create_order_date - '05JAN2013'd)/7) + 1;
  weekday = weekday(Create_order_date);
  price = sales/product_num;
```

```
proc means
  data=bysj.product min mean max p1 p5 p10 p25 p50 p75 p90 p95 p99;
  var price;
  class month;
quit;
```

初步认识数据


淘宝电商数据分析
销售额
x 10000
订单量
百
80
70
60
50
40
30
20
10
0
40
35
30
25
20
15
10
5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
2013年2月10日春节
COUNT_of_Order
SUM_of_Sales

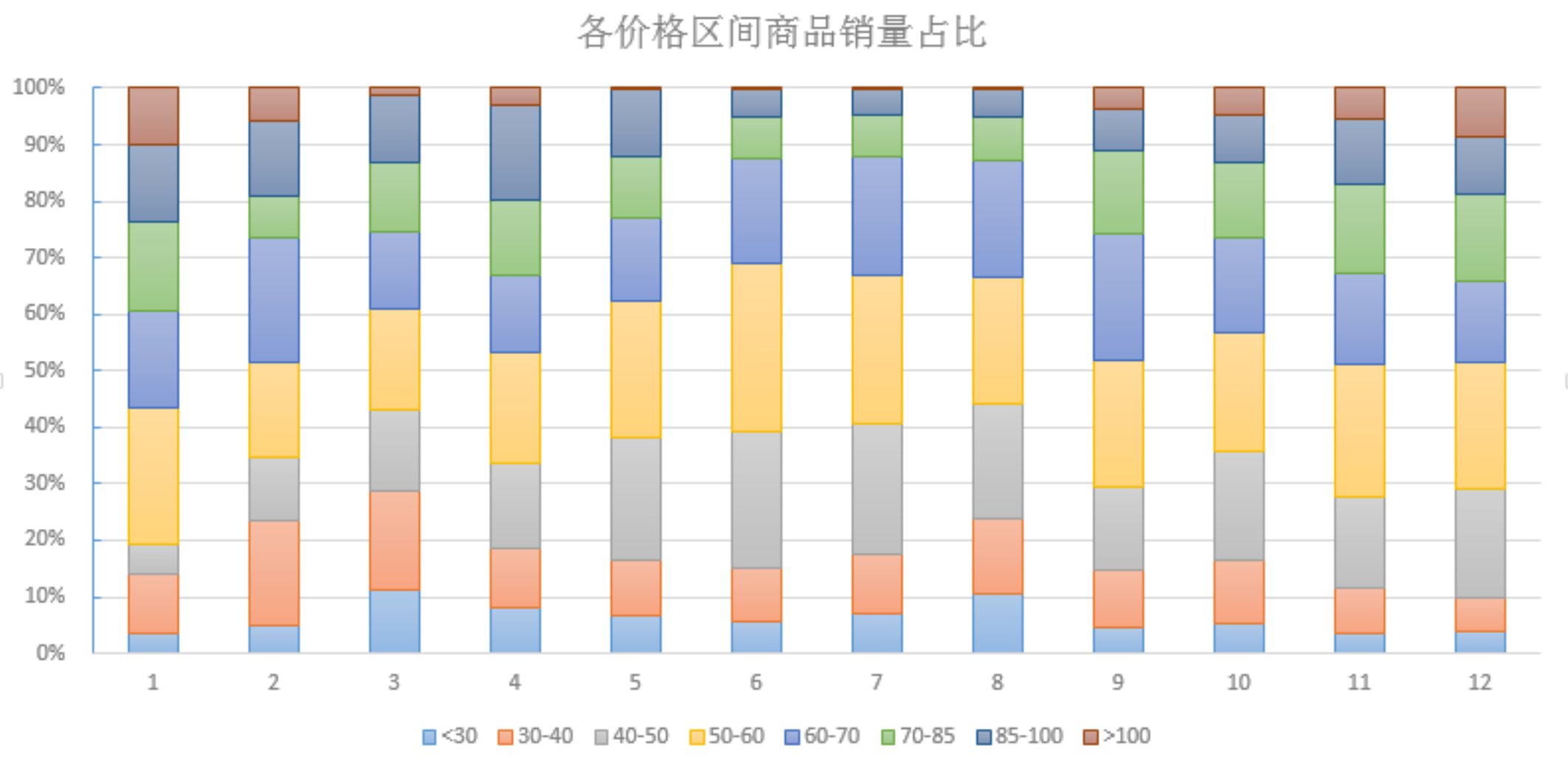
简单描述——商品分析



个百位数	第 50 个百分位数	第 75 个百分位数	第 90 个百分位数	第 95 个百分位数	第 99 个百分位数
00000	66.3600000	82.7600000	100.5250000	139.0000000	228.0000000
00000	59.3400000	73.9750000	98.0000000	116.0500000	149.0000000
00000	56.0375000	70.4520000	89.0000000	95.0000000	114.0000000
00000	59.0000000	79.0000000	95.0000000	96.3333333	119.0000000

```
/*价格分段*/
data bysj.product;
  set bysj.product;
  length price_flag $ 12;
  if price < 30 then price_flag = "<30";
  else if price < 40 then price_flag = "30-40";
  else if price < 50 then price_flag = "40-50";
  else if price < 60 then price_flag = "50-60";
  else if price < 70 then price_flag = "60-70";
  else if price < 85 then price_flag = "70-85";
  else if price < 100 then price_flag = "85-100";
  else price_flag = ">100";
  format price_flag $8.;
run;
```

简单描述——商品分析


各价格区间商品销量占比
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
8
9
10
11
12
<30
30-40
40-50
50-60
60-70
70-85
85-100
>100

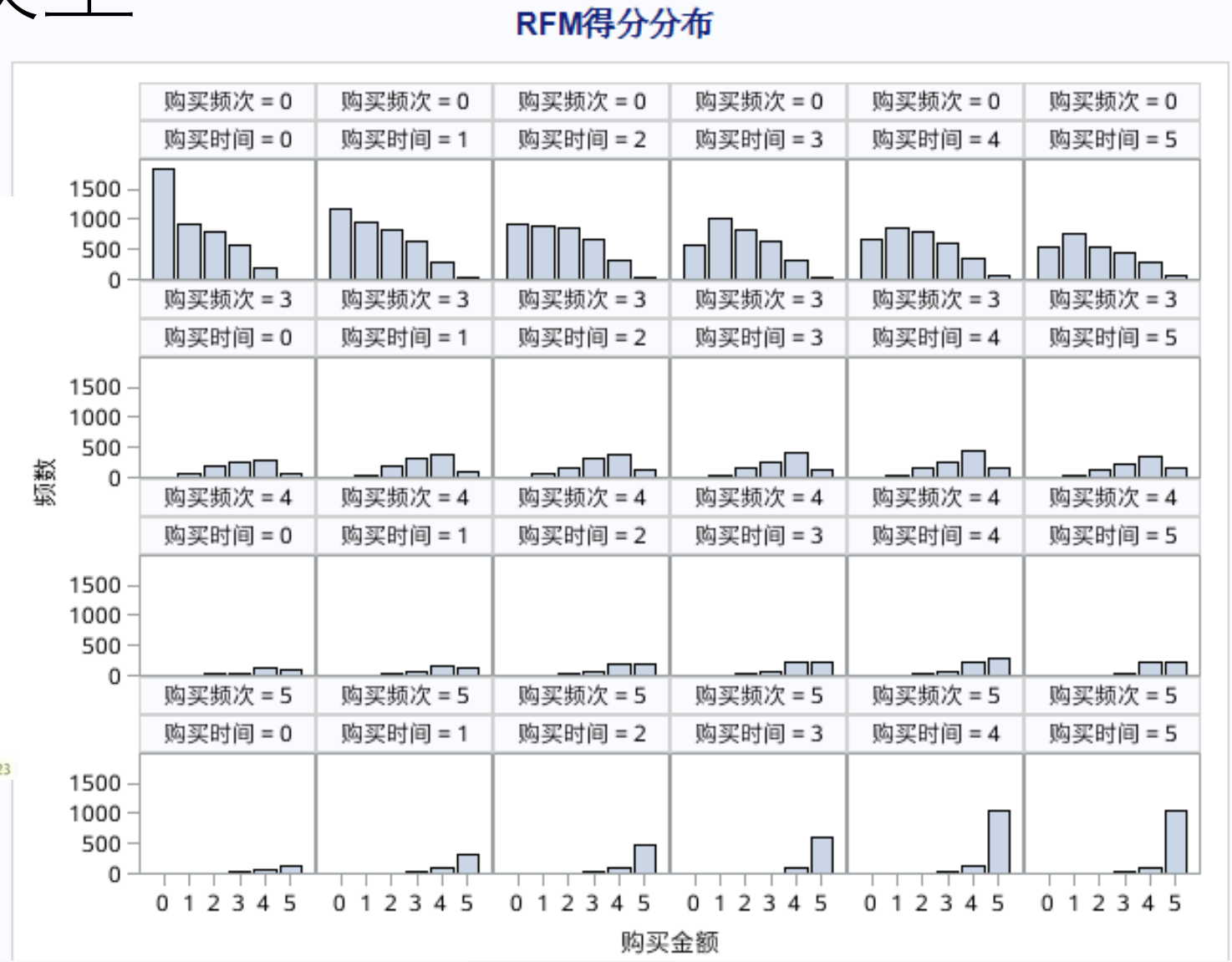
深入挖掘——客户分析

RFM客户分类模型

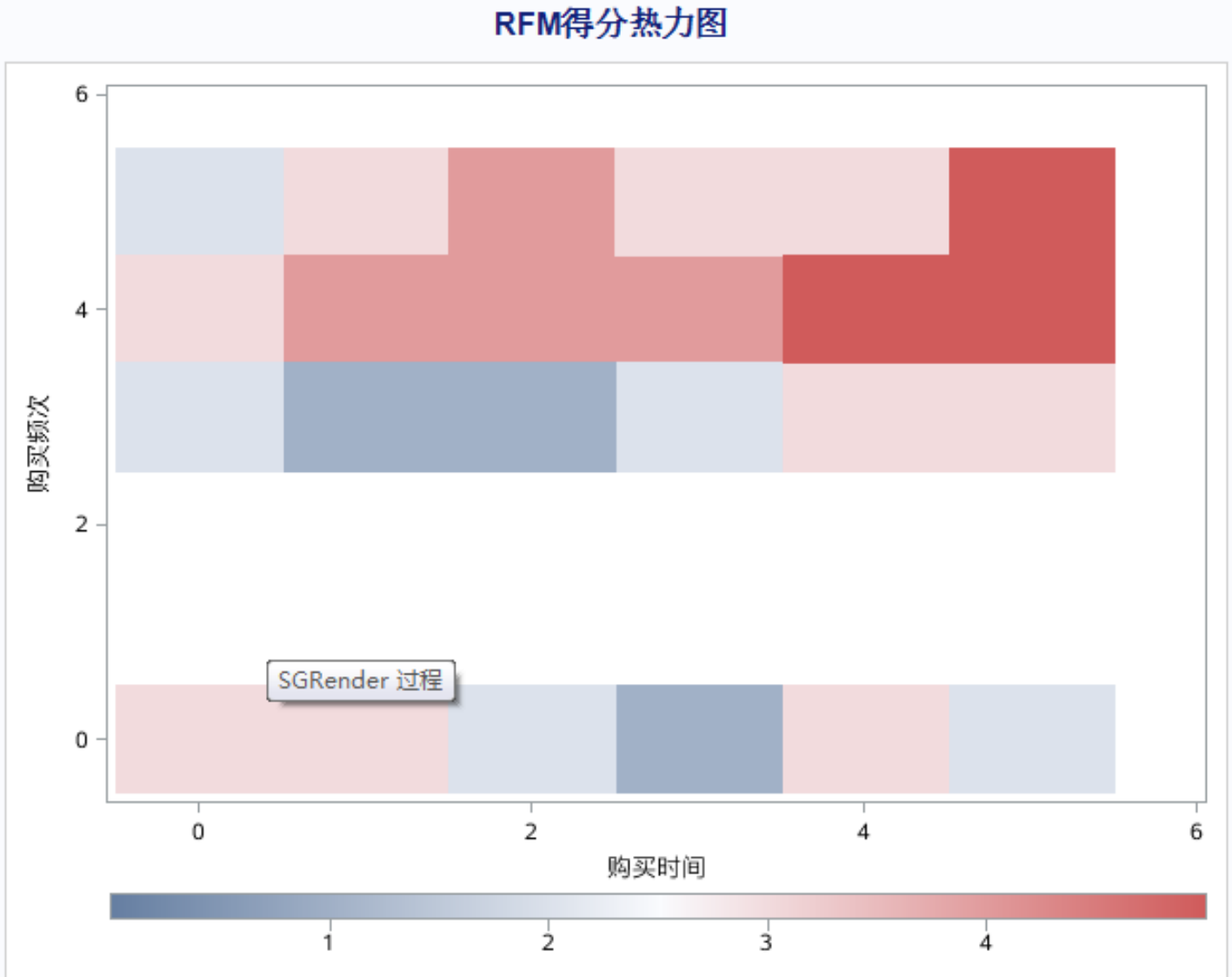
重复购买建模分析

RFM客户分类模型

R	F	M
Recency	Frequency	Money
最近一次消费间隔时间	最近一段时间购买次数	某段时间内的累积消费金额



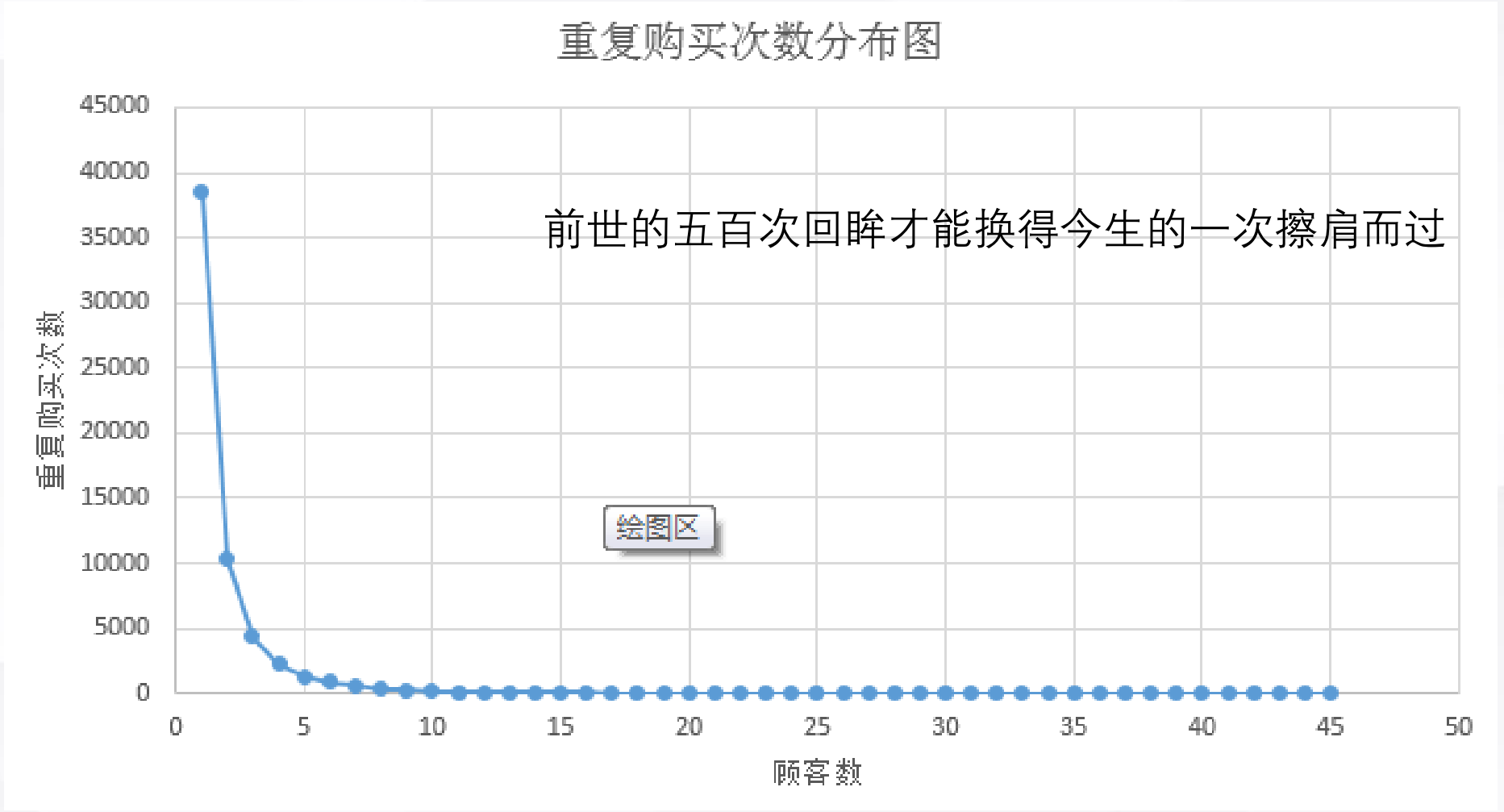
RFM客户分类模型



找准目标客户

对症下药

重复购买建模分析


重复购买次数分布图
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
重复购买次数
0
5
10
15
20
25
30
35
40
45
50
顾客数
绘图区
前世的五百次回眸才能换得今生的一次擦肩而过

重复购买建模分析

- ◆分割数据集：训练集70%，测试集30%
- ◆构建dmdb
- ◆Dmsplit方法筛选变量
- ◆建立logistic回归模型
- ◆测试模型
- ◆模型评价

The DMSPLIT Procedure

Effect Summary

最大似然估计分析

参数	自由度	估计	标准误差	Wald 卡方	Pr > 卡方
Intercep					)01
trans_e					)01
Count_					)01
Total_a					)01
timeh					)01
Pay_amount	1	0.000508	0.000208	5.9633	0.0146
timed	1	-0.2003	0.0693	8.3566	0.0038

预测概率和观测响应的关联

一致部分所占白分比	71.7	Somers D	0.441
不一致部分所占白分比	27.6	Gamma	0.444
结值白分比	0.7	Tau-a	0.211
对	1058297778	c	0.721

timeh	32	769
Pay_amount	99	91
timed	274	73

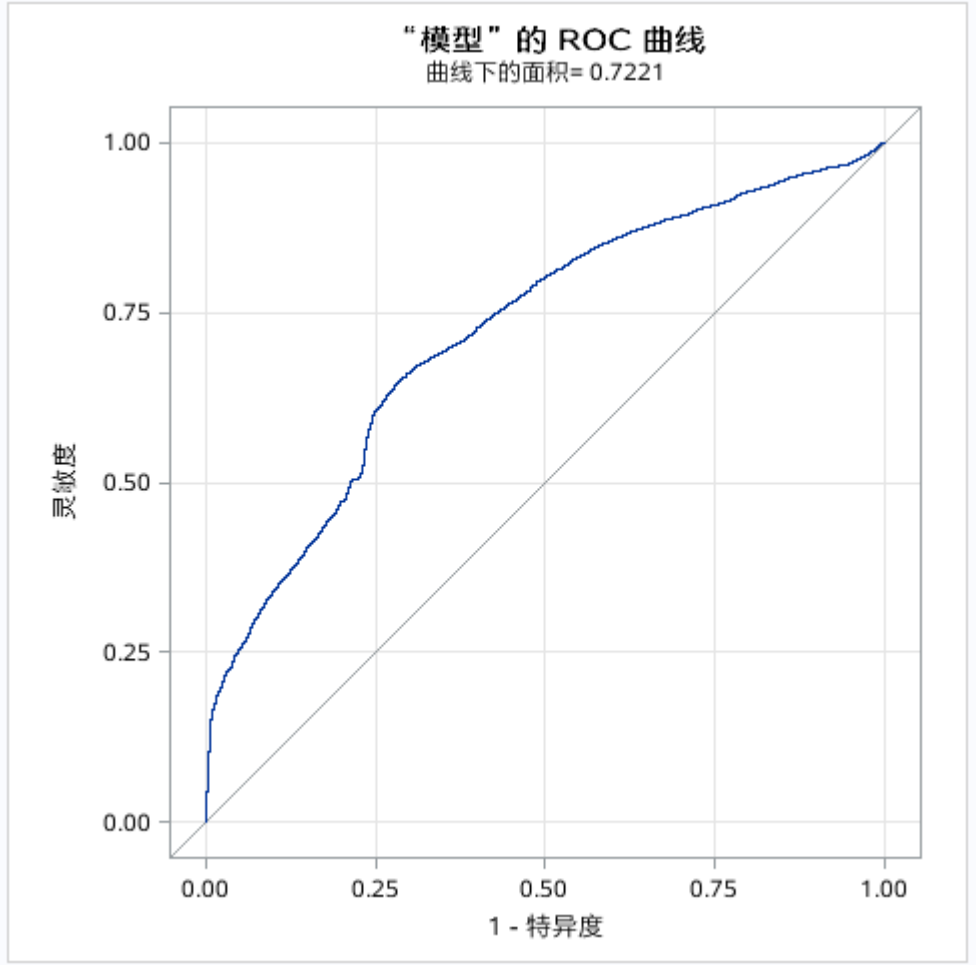
重复购买建模分析

误分类表用以判断模型优劣

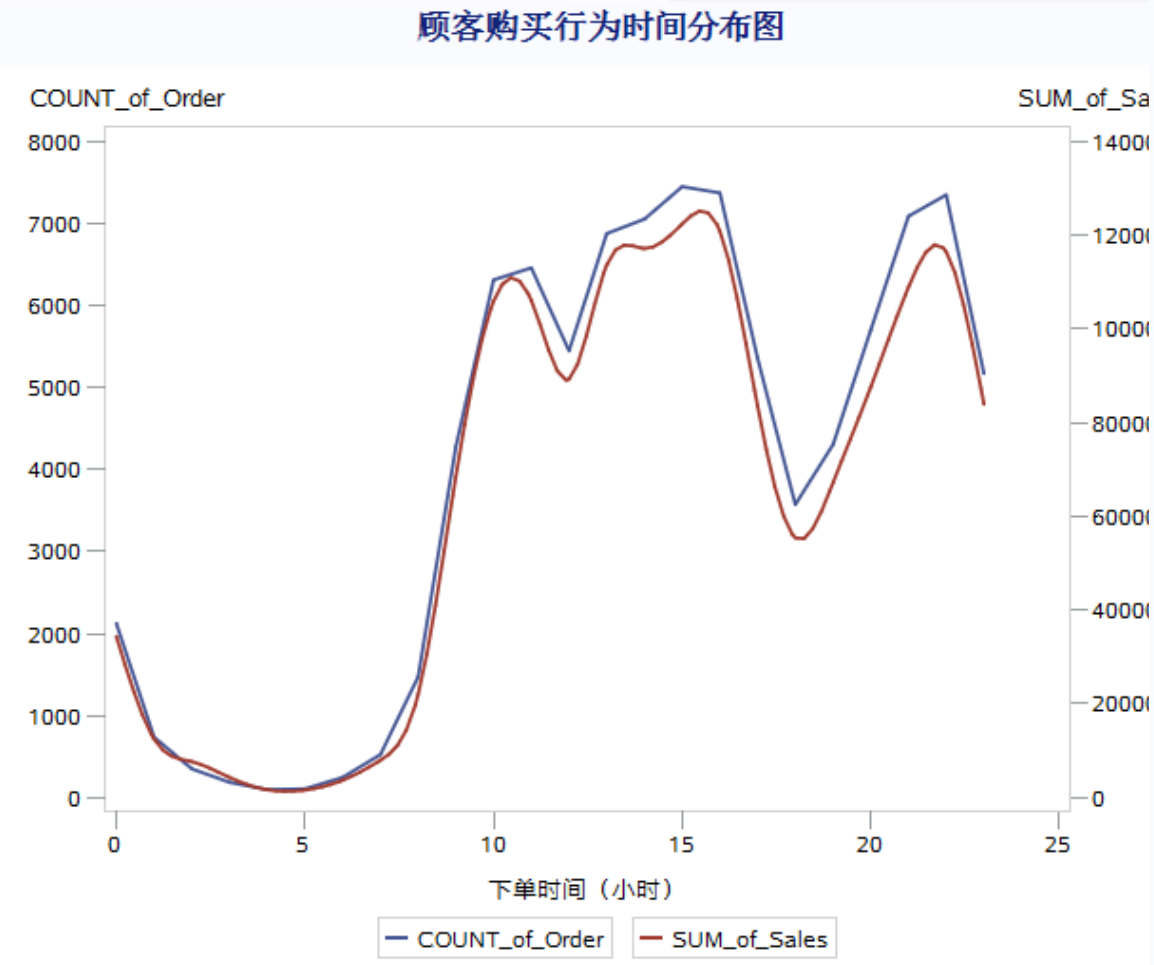
FREQ 过程

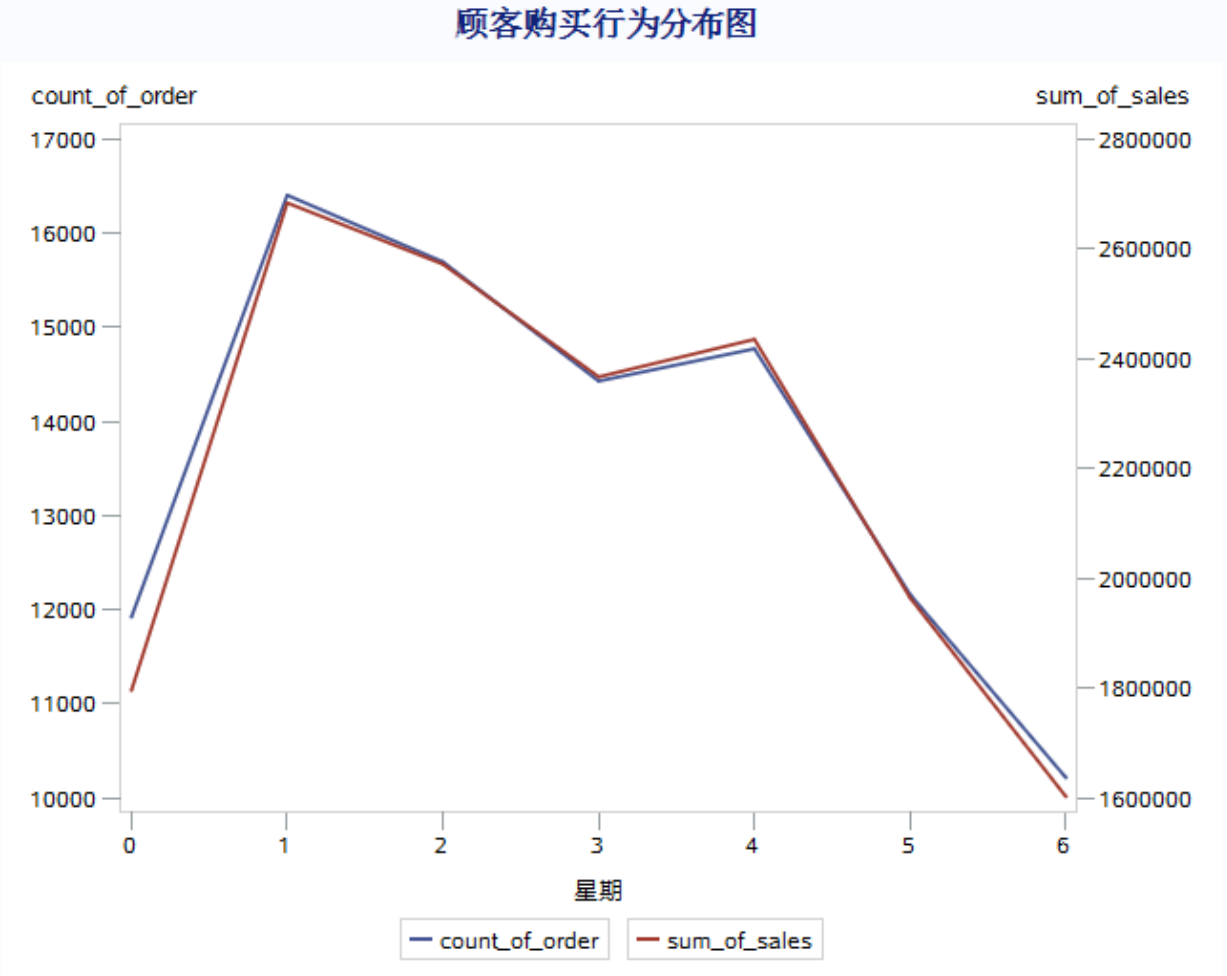
表 - F_rebuy * I_rebuy				
		I_rebuy(到: rebuy)		
		0	1	合计
F_rebuy(从: rebuy)				
0	频数	6868	4301	11169
	百分比	24.19	15.15	39.34
	行百分比	61.49	38.51	
	列百分比	58.20	25.92	
1	频数	4933	12290	17223
	百分比	17.37	43.29	60.66
	行百分比	28.64	71.36	
	列百分比	41.80	74.08	
合计	频数	11801	16591	28392
	百分比	41.56	58.44	100.00
频数缺失 = 176				

Pprob=0.45

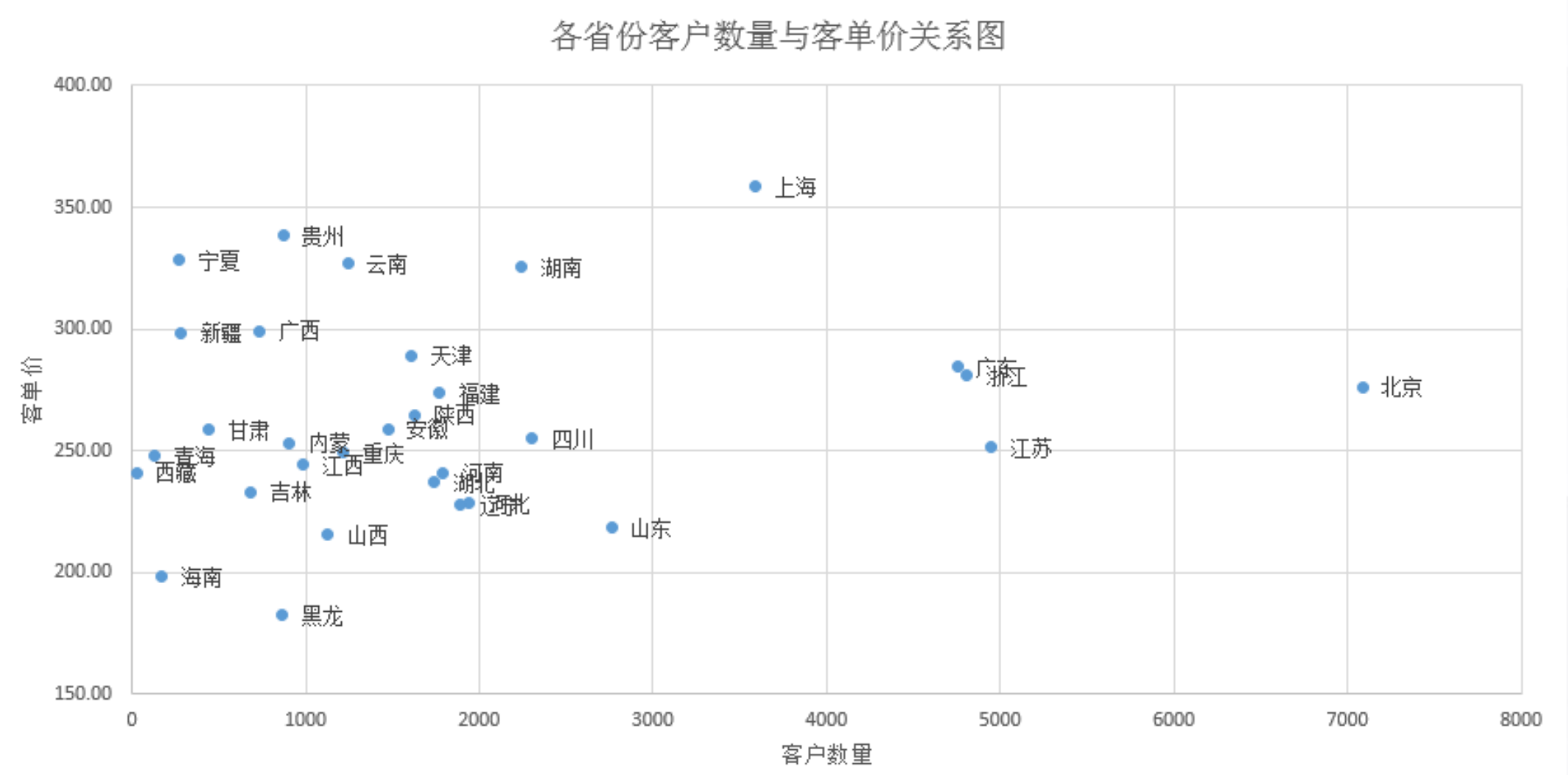


用户行为分析


顾客购买行为时间分布图
COUNT_of_Order
SUM_of_Sa
8000
7000
6000
5000
4000
3000
2000
1000
0
14000
12000
10000
8000
6000
4000
2000
0
0
5
10
15
20
25
下单时间（小时）
COUNT_of_Order
SUM_of_Sales


顾客购买行为分布图
count_of_order
sum_of_sales
17000
16000
15000
14000
13000
12000
11000
10000
2800000
2600000
2400000
2200000
2000000
1800000
1600000
0
1
2
3
4
5
6
星期
count_of_order
sum_of_sales

用户行为分析


各省份客户数量与客单价关系图
客单价
客户数量
400.00
350.00
300.00
250.00
200.00
150.00
0
1000
2000
3000
4000
5000
6000
7000
8000
上海
贵州
宁夏
云南
湖南
新疆
广西
天津
福建
陕西
甘肃
安徽
四川
内蒙
重庆
青海
江西
西藏
河南
吉林
湖北
辽宁
河北
山东
山西
海南
黑龙
广东
浙江
北京
江苏

相关建议

产品方面

客户关怀方面

客户服务方面

Q&A

THANKS