英国电商用户数据分析

E-Commerce Data Analysis

利用Python进行可视化及数据分析

项目内容

CONTENTS

数据获取与预处理

Data Acquisition and Pre-Processing

数据集简介

数据来源：

Kaggle: E-Commerce Data https://www.kaggle.com/carrie1/ecommerce-data

该数据集为英国在线零售商在2010年12月1日至2011年12月9日间发生的所有网络交易订单信息，包括客户编号、订单编号、商品代码及数量、单价等字段。

数据集内容

数据集为CSV格式，大小为46.2MB，包含8个字段、541908条数据。具体字段为：

- InvoiceNo：发票编号。为每笔订单唯一分配的6位整数。若以字母'C'开头，则表示该订单被取消。
- StockCode：产品代码。为每个产品唯一分配的编码。
- Description：产品描述。
- Quantity：数量。每笔订单中各产品分别的数量。
- InvoiceDate：发票日期和时间。每笔订单发生的日期和时间。
- UnitPrice：单价。单位产品价格，单位为英镑。
- CustomerID：客户编号。为每个客户唯一分配的5位整数。
- Country：国家。客户所在国家/地区的名称。

数据预处理

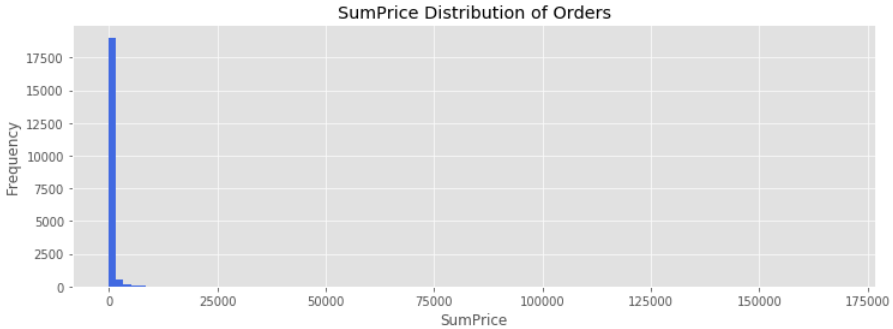
- 导入数据
- 重复值处理
- 缺失值处理
- 时间、日期处理
- 异常值处理：被取消订单、单价为负的订单、单价为0的订单
- 新增字段

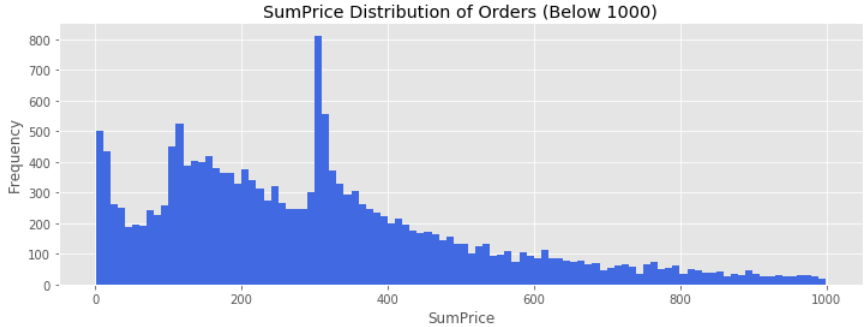
数据可视化与用户分析

Data Visualization and User Analysis

订单金额分布

	Quantity	SumPrice
count	19960.000000	19960.000000
mean	279.179359	533.171884
std	955.011810	1780.412288
min	1.000000	0.380000
25%	69.000000	151.695000
50%	150.000000	303.300000
75%	296.000000	493.462500
max	80995.000000	168469.600000

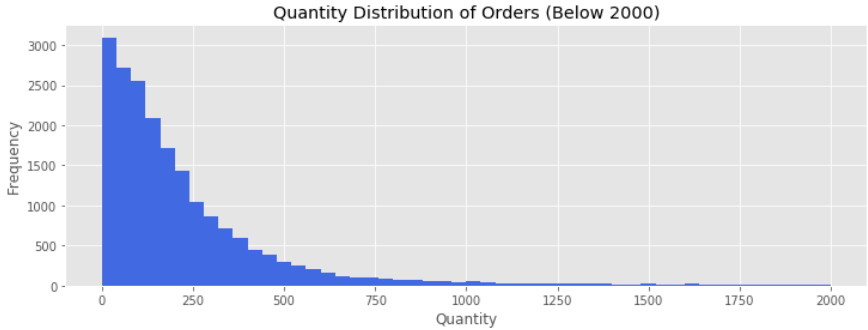




从2010.12.01到2011.12.09，共产生有效订单19960笔，订单交易金额的平均数远高于中位数，甚至高于Q3分位数，说明订单的总体差异大，存在金额极大的订单/购买力极强的客户。订单金额集中在0-20英镑，100-200英镑和300-320英镑。

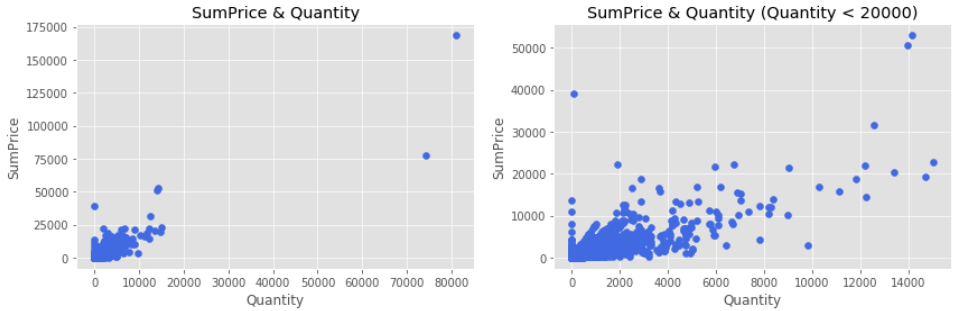
订单内商品数量分布

	Quantity	SumPrice
count	19960.000000	19960.000000
mean	279.179359	533.171884
std	955.011810	1780.412288
min	1.000000	0.380000
25%	69.000000	151.695000
50%	150.000000	303.300000
75%	296.000000	493.462500
max	80995.000000	168469.600000



订单内的商品数量呈现出典型的长尾分布，大部分订单的商品数量在250件内，商品数量越多，订单数相对越少。

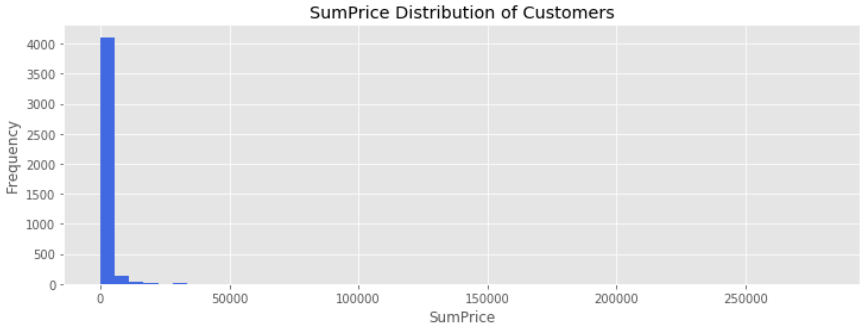
订单金额 vs 订单内商品数量

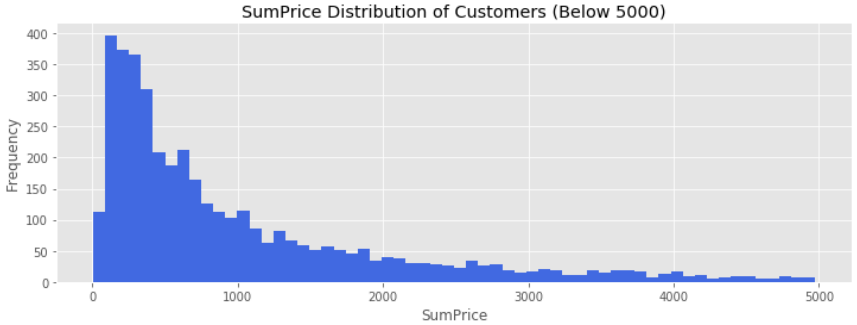


总体来说订单交易金额与订单内商品件数是正相关的，订单内的商品数越多，订单金额也相对越高。但在Quantity靠近0的位置也有若干量少高价的订单，后续可以试探究。

客户消费金额分布

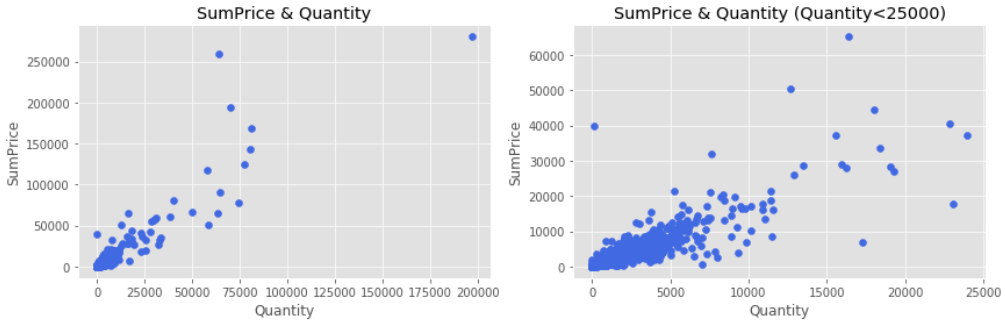
	InvoiceNo	Quantity	SumPrice
count	4338.000000	4338.000000	4338.000000
mean	4.272015	1187.644537	2048.688081
std	7.697998	5043.619654	8985.230220
min	1.000000	1.000000	3.750000
25%	1.000000	159.000000	306.482500
50%	2.000000	378.000000	668.570000
75%	5.000000	989.750000	1660.597500
max	209.000000	196915.000000	280206.020000





从直方图看，大部分用户的消费能力不高，高消费用户在图上几乎看不到。这也确实符合消费行为的行业规律。与前面订单金额的多峰分布相比，客户消费金额的分布呈现单峰长尾形态，金额大约集中在80-1000英镑间。

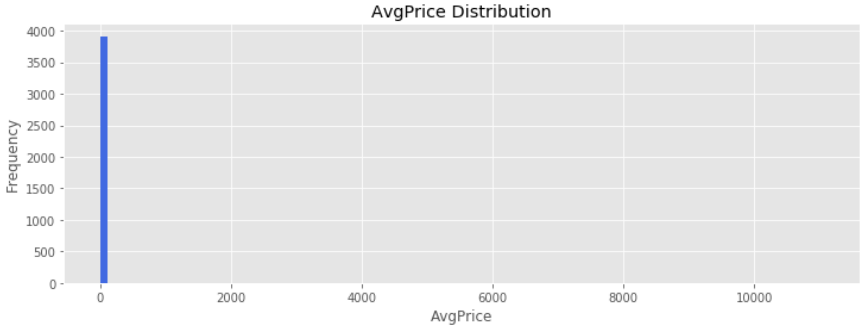
客户消费金额 vs 消费商品件数

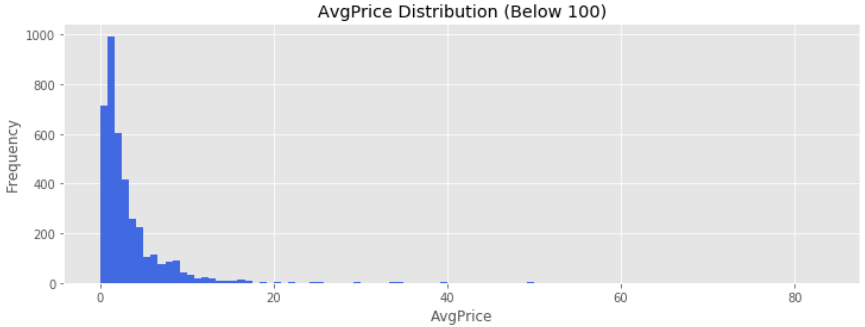


客户群体比较健康，而且规律性比订单更强，同时拥有一定数量消费能力强的用户。总体来说客户的消费金额与购买的商品数量是正相关的，客户购买的东西越多，消费金额相对就越高。

商品的平均价格分布

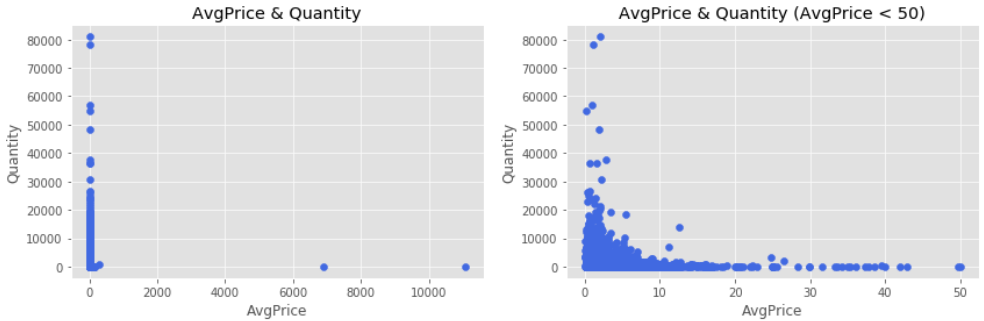
StockCode	Quantity	SumPrice	AvgPrice
10002	860	759.89	0.883593
10080	303	119.09	0.393036
10120	192	40.32	0.210000
10123C	5	3.25	0.650000
10124A	16	6.72	0.420000





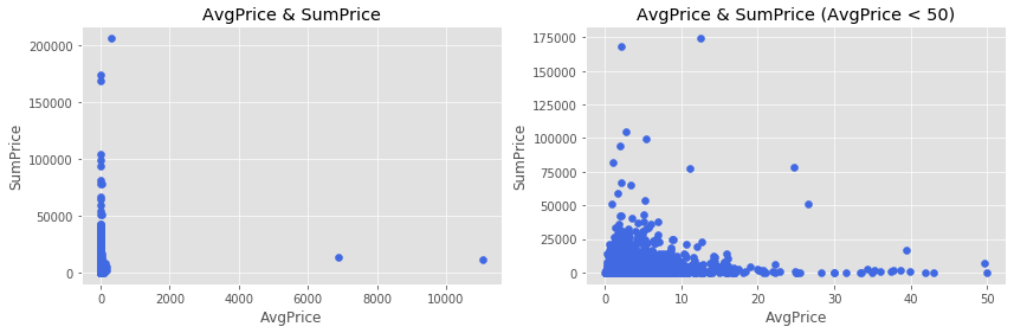
商品的平均价格基本集中在100英镑以内，峰值是1-2英镑，单价10英镑以上的商品已经很少见，看来该电商的定位主要是价格低的小商品市场。

商品销售量 vs 商品平均价格



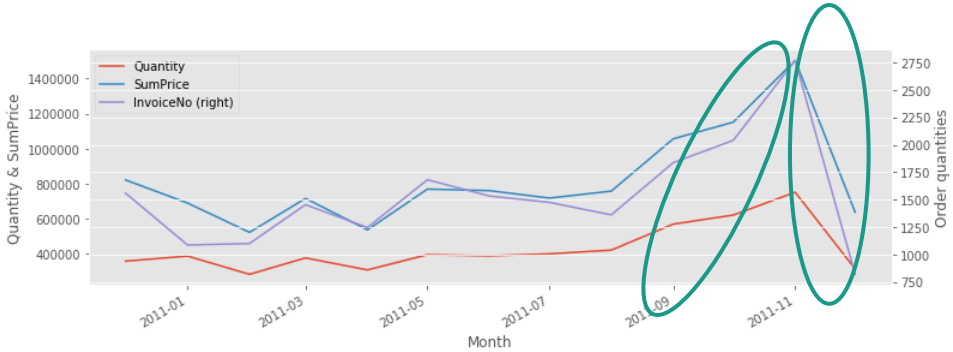
从商品的销量上来看，毫无疑问是低于5英镑的低价区商品大获全胜，受到了客户们的喜爱。

销售额 vs 商品平均价格



低价区的商品除了在销售数量占据绝大多数，也同时构成了销售额的主要部分；高价的商品虽然单价高昂，但销量很低，并没有带来太多的销售额。

销售额、销售量、订单数量 vs 月份



需要注意此处2011年12月仅统计了前9天，如果全月能基本保持前9天的销售情况，销售额会远超2010年同期。

三条折线总体上呈现相近的趋势，除了2011年2月和4月略低外，2010年12月至2011年8月基本维持相近的销售情况；随后在9月-11月连续增长，达到高峰。考虑因素：万圣节(11.1)，圣诞节(12.25)都在年末，与图中年末趋势相呼应。

用户生命周期分布

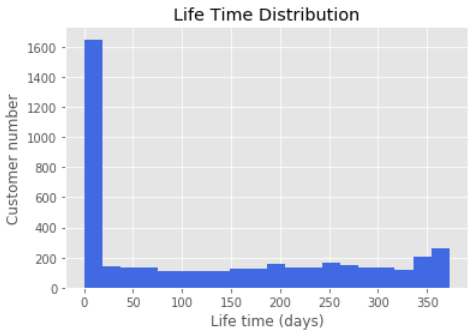
初次消费时间	Count
2010-12-01	95
2010-12-02	93
2010-12-08	83
2010-12-06	70
2010-12-05	69
2010-12-09	67
2010-12-16	58
2010-12-07	50
2010-12-03	46
2010-12-15	42

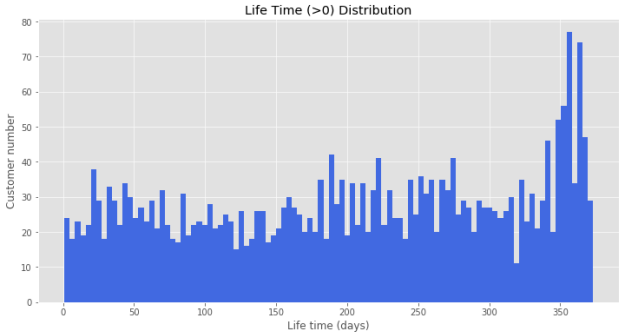
末次消费时间	Count
2011-12-08	103
2011-12-06	94
2011-12-05	94
2011-12-07	90
2011-12-01	79
2011-11-29	77
2011-11-22	74
2011-12-02	72
2011-11-30	71
2011-11-17	64

	用户生命周期
count	4338
mean	130 days
std	132 days
min	0 days
25%	0 days
50%	93 days
75%	252 days
max	373 days

发现初次消费的高频日期为统计时段的初期，末次消费的高频日期为统计时段的末期。说明有大量用户的生命周期被低估，实际上还要向前向后延伸。共有4338个有CustomerID的客户，其平均生命周期为130天，中位数则是93天，说明有部分生命周期很长的忠实客户拉高了均值；而最小值和Q1分位数都为0天，说明存在25%以上的客户仅消费了一次，生命周期的分布呈两极分化的状态。

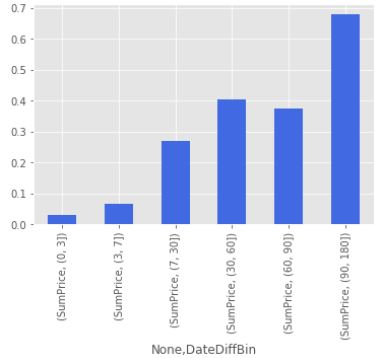
用户生命周期分布





将生命周期为0的客户去除后，我们发现生命周期在0-75天的客户数略高于75-170天，约1/4的客户集中在170天-330天，属于较高质量客户的生命周期；而在330天以后，则是数量可观的长期客户，拥有极高的用户粘性。考虑到这些客户中有许多未进行完整的生命周期，实际的客户平均生命周期会更长。

用户留存分布

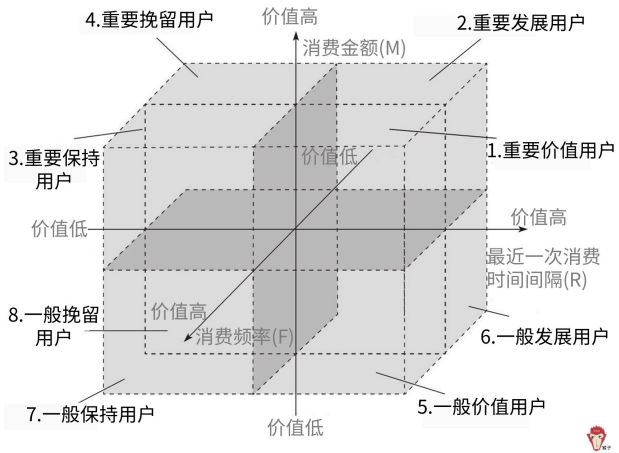


我们将用户的消费时间间距按照3、7、30、60、90、180天分组，去除消费间距为0的客户(即只有过一次购买行为的客户)，发现：在这些老客户中，只有3.2%在第一次消费的次日至3天内有过消费，6.6%的客户在4-7天有过消费。分别有40.5%和37.4%的客户在首次消费后的第二个月内和第三个月内有过购买行为。将时间范围继续放宽，有高达67%的客户在90天至半年内消费过。说明该电商网站的客户群体，其采购并非高频行为，但留存下来的老客户忠诚度却极高。

用RFM模型进行用户分类

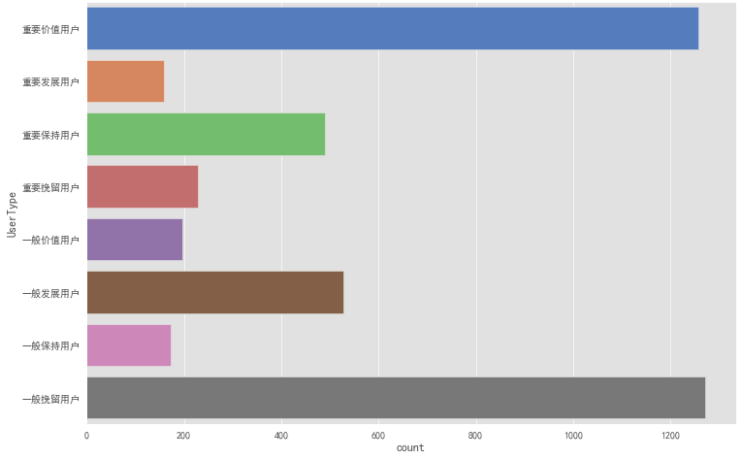
Customer Classification Using RFM

RFM模型



	CustomerID	Recency	Frequency	MonetaryValue	R_Score	F_Score	M_Score	UserType
0	17850	372	297	5391.21	low	high	high	重要保持用户
1	13047	31	172	3237.54	high	high	high	重要价值用户
2	12583	2	247	7281.38	high	high	high	重要价值用户
3	13748	95	28	948.25	low	low	high	重要挽留用户
4	15100	333	3	876.00	low	low	high	重要挽留用户
...	...	...	...	...	...	...	...	...
4332	15471	2	73	454.48	high	high	low	一般价值用户
4333	13436	1	12	196.89	high	low	low	一般发展用户
4334	15520	1	18	343.50	high	low	low	一般发展用户
4335	13298	1	2	360.00	high	low	low	一般发展用户
4336	14569	1	12	227.39	high	low	low	一般发展用户

RFM模型



我们发现在此电商的客户群中，客户分类呈两极分化的状态，重要价值用户和一般挽留用户占据了超过58%的比例。对于重要价值用户，属于优质用户，需要注意保持。一般挽留用户，建议更加重视客户初次消费的体验感，可以考虑通过调查问卷、服务评分等方式获知新客对于购买流程中不满意之处，针对性地加以改进；并且花更多的精力引导其进行再次消费，如发放有时限的优惠券等。

启发与结论

Insights and Conclusions

1.订单维度：

有效订单共19960笔，笔单价为533.17英镑，连带率约为279件。订单以批发性质为主，订单间差异较大，存在部分购买力极强的客户。总体来说订单交易金额与订单内商品件数正相关。

2.客户维度：

客单价为2049英镑，客户的购买力存在较大差距，拥有一定数量消费能力强的客户。客户群体比较健康，其消费金额与购买商品数量正相关，而且规律性比订单更强。

3.商品维度：

商品的单价会发生波动，集中于1-2英镑，定位主要是低价的小商品市场。低于5英镑的商品最受客户喜爱，同时也构成了销售额的主要部分。高价的商品虽然单价不菲，但销量很低，并没有带来太多的销售额。建议主要针对销售价格低于10英镑的产品，来进一步扩充低价区的品类。

4.时间维度：

订单数、销量、销售额总体上呈现相近的趋势， 在2011年9月-11月连续增长，达到高峰。考虑到主营商品为礼品，猜测受节日影响可能较大，如万圣节（11月1日）和圣诞节（12月25日）影响。

5.生命周期：

客户平均生命周期为130天，生命周期的分布呈两极分化的状态。消费两次及以上的客户平均生命周期是203天，远高于总体均值103天。建议更加重视客户初次消费的体验感，可以考虑通过调查问卷、服务评分等方式获知新客对于购买流程中不满意之处，针对性地加以改进；并且花更多的精力引导其进行再次消费，如发放有时限的优惠券等。

6.留存情况：

客户群体的采购并非高频行为，但留存下来的老客户忠诚度极高。而仅有首次购买行为的客户占总客户的37.5%，如能提高这部分群体的留存率，将会带来很高的收益。