# 阿里云池淘宝用户行为数据分析

### 原始数据的处理

本数据集来源于阿里天池，包含了淘宝App由2017年11月25日至2017年12月3日之间，有行为的随机用户的所有行为（包括点击、购买、加购、收藏）。

数据集的组织形式即每一行表示一条用户行为。

由于数据量过大，约有1亿条用户行为，所以抽样约10万的用户行为进行分析。

create table userbehavior

as

select \* from 原始数据文件

limit 100000;



1. **选择子集**

数据集中有5个字段，在后续分析中都需要用到，不需要删除，保留即可。

1. **列名重命名**

原数据无列名，在导入数据添加字段时，已将列名命好，这里不需要重新命名了。

1. **查看重复值**

由于该数据集无主键，故需要对整行重复的内容进行处理。

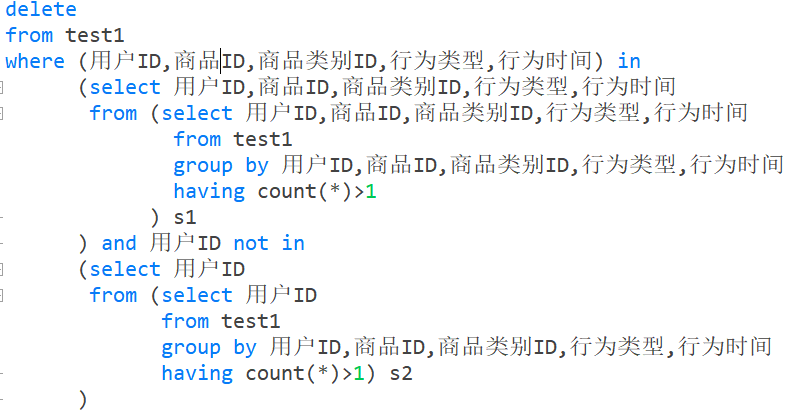
select \*

from userbehavior

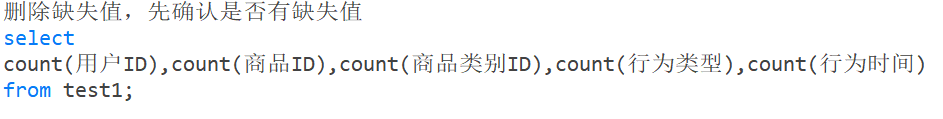
group by 用户ID,商品ID,商品类别ID,行为类型,行为时间

having count(\*)>1;

1. **全字段去重**

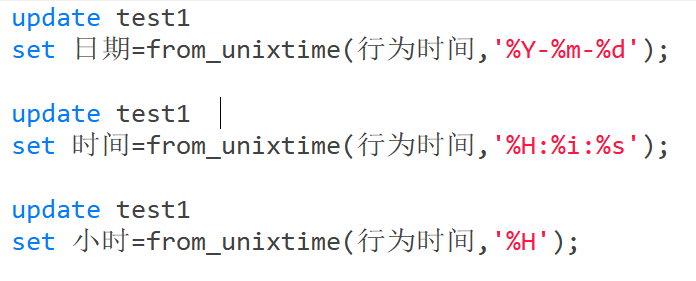


1. **缺失化处理**



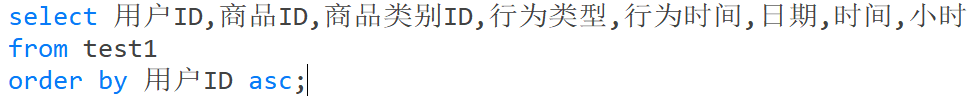
1. **一致性处理**

一致性的检验与字段格式的说明有关，这里主要检查数据的写入格式是否符合统一规范。此时timestamps这一列数据为unixtime，需要利用from\_unixtime函数转换为：日期格式、时间格式、小时数。



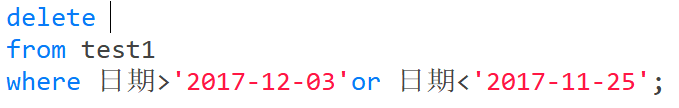
1. **数据排序**

对数据按照用户ID顺序，进行降序排序。



1. **异常值处理**

因为原数据集(UserBehavior.csv)包含的数据是2017年11月25日至2017年12月3日之间的数据，若出现了不在这时间段的数据，则为异常数据。



### 一、行为分析

基于群组分析方法，研究用户活跃度以及用户的购买情况。

* + - 1. 总揽独立访客量，商品总数量，商品类型总数等

select count(distinct 用户ID) as 独立访客量,

count(distinct 商品ID) as 商品总数量,

count(distinct 商品类别ID) as 商品类型总数,

count(distinct 行为类型) as 行为类型总数,

count(distinct 日期) as 天数,

min(日期) 最早日期,

max(日期) 最晚日期

from test1;



清洗后的数据记录了2017年11月25日至2017年12月3日期间，独立访客数即用户数为983人，商品总数为64439，商品类型有3128，用户行为类型有4种：点击、加购、收藏、购买，天数为9天。

* + - 1. 查看总体UV，PV，人均浏览次数，成交量

pv总数

人均浏览次数=pv总数/独立访客数

成交量总数

select count(distinct 用户ID) as 独立访客数,

sum(case when 行为类型='pv' then 1 else 0 end) as 点击数,

sum(case when 行为类型='pv' then 1 else 0 end)/count(distinct 用户ID) as 人均浏览次数,

sum(case when 行为类型='buy' then 1 else 0 end) as 成交量

from test1;



独立访客数983人，点击数89664人，人均浏览次数91.2146，成交量2101

* + - 1. 查看日均UV,PV，人均浏览次数，成交量

select 日期, count(distinct 用户ID) as 独立访客数,

sum(case when 行为类型='pv' then 1 else 0 end) as 点击数,

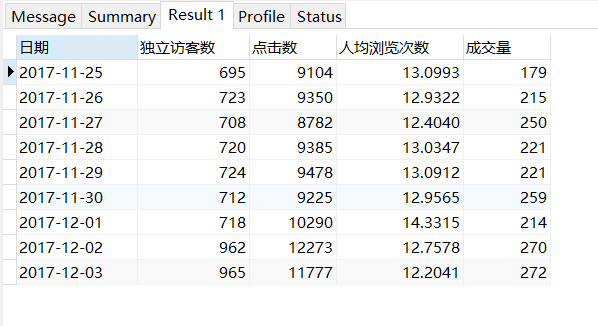
sum(case when 行为类型='pv' then 1 else 0 end)/count(distinct 用户ID) as 人均浏览次数,

sum(case when 行为类型='buy' then 1 else 0 end) as 成交量

from test1

group by 日期

order by 日期



从日均独立访客量、人均浏览次数、成交量可以看出，自12.2日起有大幅增加，推测可能由于双十二活动预热导致比前一个周末（11月27-28日）要上涨很多。



* + - 1. 查看时均UV,PV，人均浏览次数，成交量

select 小时, count(distinct 用户ID) as 独立访客数,

sum(case when 行为类型='pv' then 1 else 0 end) as 点击数,

sum(case when 行为类型='pv' then 1 else 0 end)/count(distinct 用户ID) as 人均浏览次数,

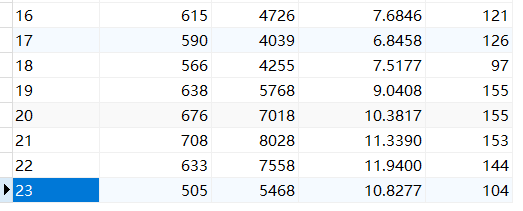
sum(case when 行为类型='buy' then 1 else 0 end) as 成交量

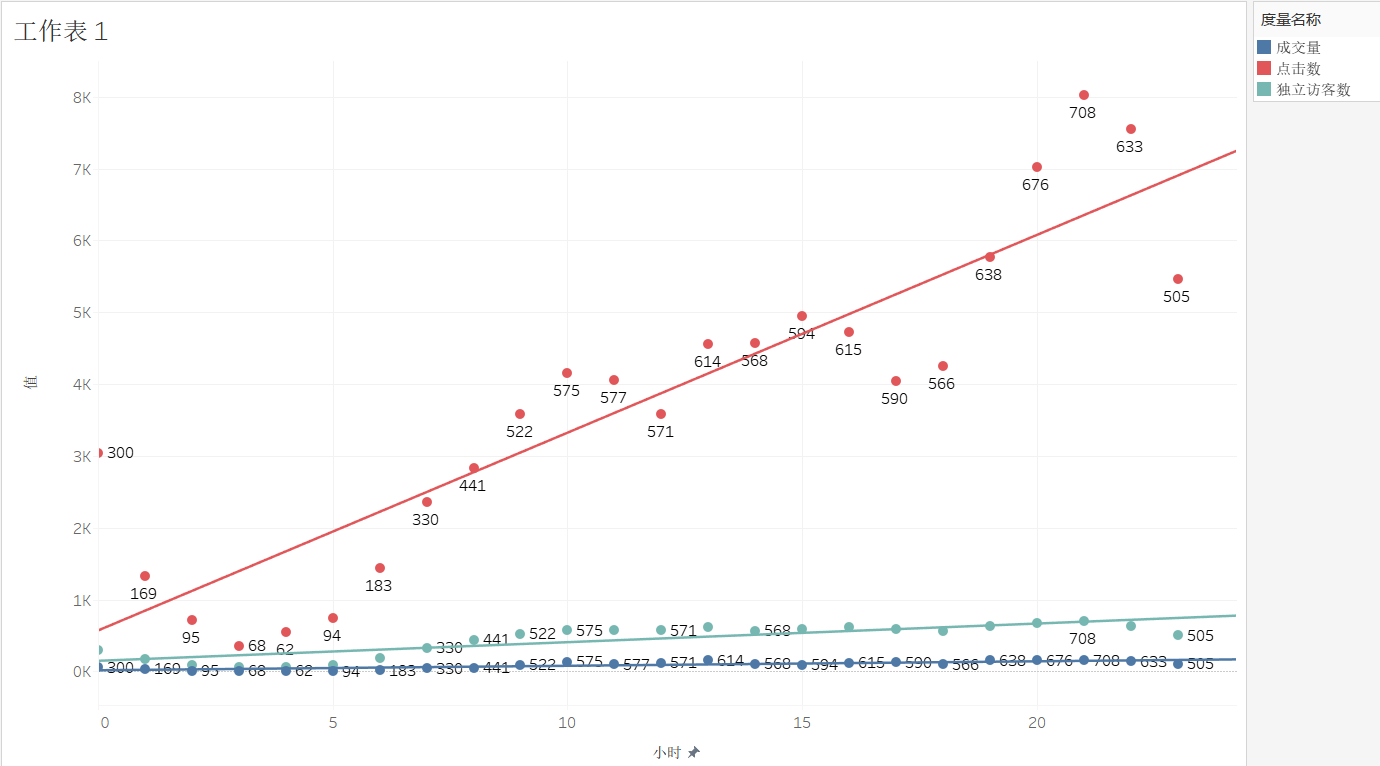
from test1

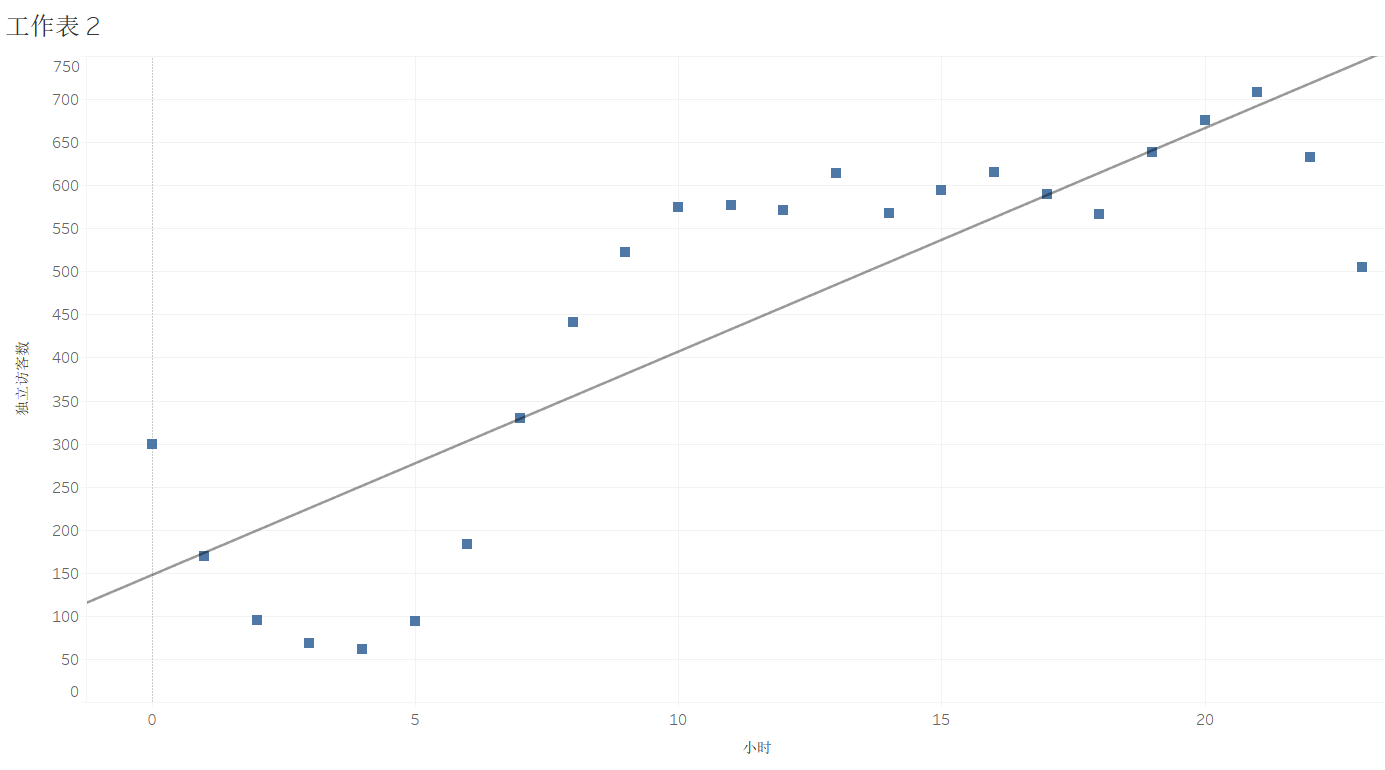
group by 小时

order by 小时;









通过以上数据及可视化图表可知，独立访客量和成交量在中午12时和晚上21-22时，达到顶峰，是用户最活跃的时刻。商家应该集中资源，重点在这两个时间段进行引流与营销活动。

* + - 1. 用户行为数据整理

查看用户点击数，加购数，收藏量，成交量

create view 用户行为数据 as

select 用户ID, count(行为类型) as 用户行为数,

sum(case when 行为类型='pv' then 1 else 0 end) as 点击数,

sum(case when 行为类型='cart' then 1 else 0 end) as 加购数,

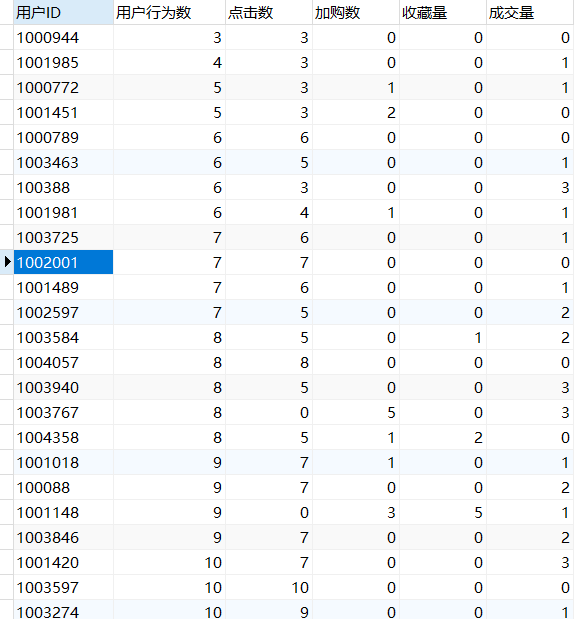
sum(case when 行为类型='fav' then 1 else 0 end) as 收藏量,

sum(case when 行为类型='buy' then 1 else 0 end) as 成交量

from test1

group by 用户ID

order by 用户行为数;



### 环节分析:

基于漏斗分析方法研究用户群体在整个购物过程中的转化/流失情况并提出如何改善转化率.

查看独立访客量，点击量，加购数，收藏量，成交量，人均点击次数



由上图可知数据总概况：独立访客数为983，用户共点击了89664次，加购了5446次，收藏了2744次，最后共成交2101次，每个独立用户平均点击了91.2146次，购买转化率为68.26%。

众所周知，在淘宝购物时，用户行为路径可分为四部分：点击-加入购物车-收藏-购买，以下查看用户在四个环节的点击量

select 行为类型,

count(distinct 用户ID) AS '独立访客量',

sum(case when 行为类型='pv' then 1 else 0 end) as '点击',

sum(case when 行为类型='cart' then 1 else 0 end) as '加购',

sum(case when 行为类型='fav' then 1 else 0 end) as '收藏',

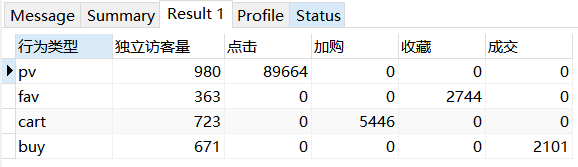
sum(case when 行为类型 ='buy' then 1 else 0 end) as '成交'

from test1

group by 行为类型

order by 行为类型 desc;

结果如下图：

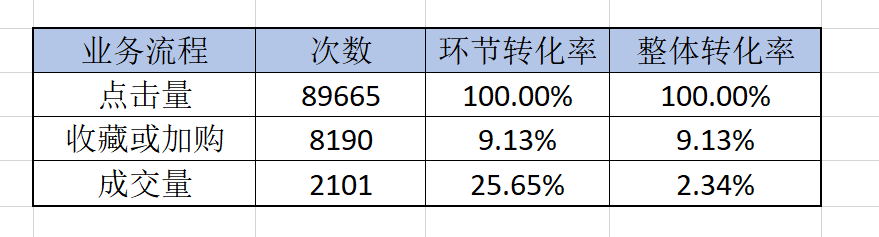


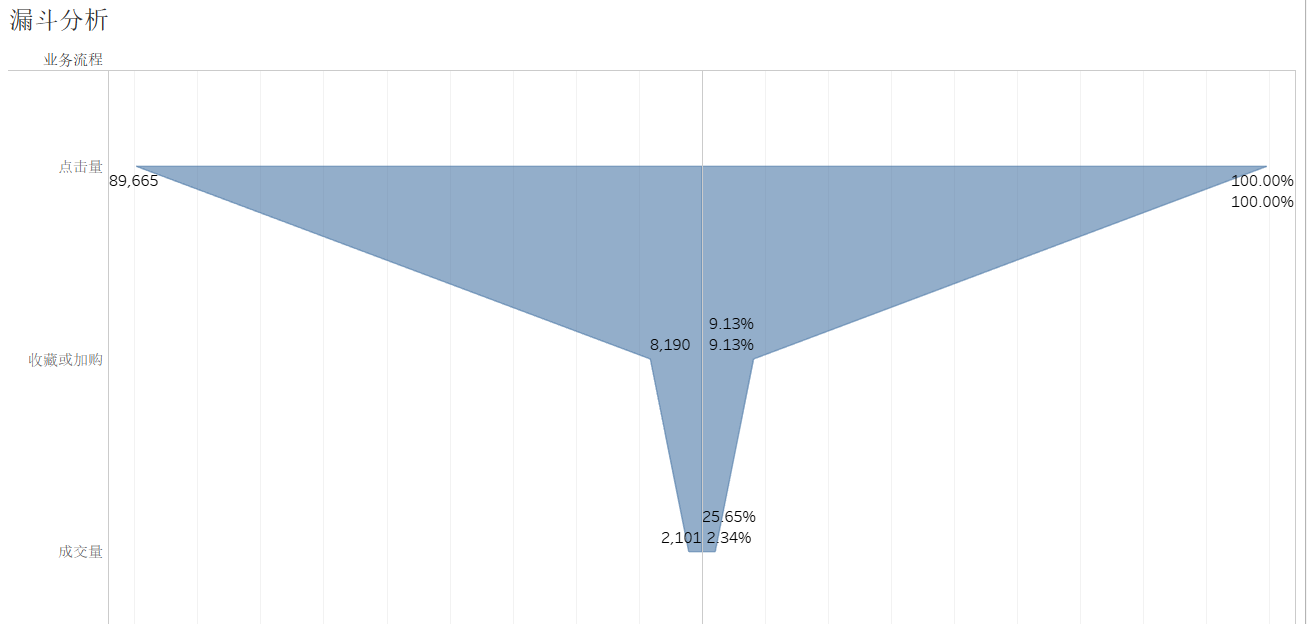
如图，得到独立用户的行为数据，980个用户总共进行了89664次点击，363个用户共进行了2744次收藏，723个用户共进行了5446次加购，而最终只有671个用户进行了2101次购买。

为了更好地运用漏斗分析方法，这里还需要计算各环节转化率：

环节转化率=本环节用户数/上一环节用户数

整体转化率=某环节用户数/第一环节用户数。





如图可知，980个用户总共进行了89665次点击浏览后，收藏或加购的次数却变成了8190，从点击-收藏或加购的转化率变成了9.13%，这意味着用户点击浏览后，却不愿意将商品收藏或加入购物车了，用户流失主要发生在点击-加购或收藏这一环节。

我们猜测发生这种情况的原因，可能是因为用户点击加入收藏或加购的过程太过于复杂，导致用户不愿继续进行下一步操作导致的，这时我们可以对App流程进行优化，使得用户点击这部分环节的操作更简捷。当然，也有可能是因为宣传价格和实际加购价格差异导致的，这样可以优化规范一下平台规则。

### 商品分析

找出复购率为Top10的用户所购买的Top10的商品，以及点击量,收藏量,加购量及购买量Top10的商品

1. **从用户角度上**

整体统计有购买行为的用户总数

select count(distinct 用户ID) as 购买总人数

from userbehavior

where 行为类型='buy';

分别按照购买次数统计出总人数

select 购买次数,count(\*) as 人数

from(select 用户ID,count(用户ID) as 购买次数

from test1

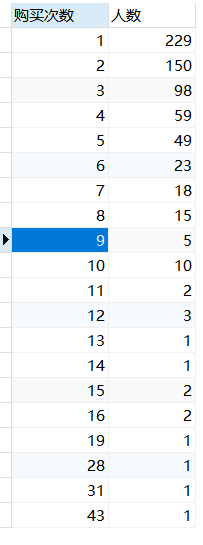
where 行为类型='buy'

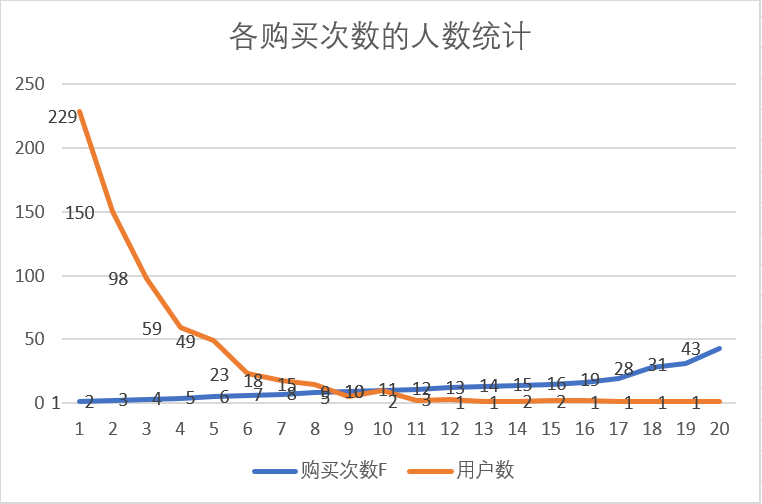
group by 用户ID

having count(用户ID)>=1) as 用户购买

group by 购买次数

order by 购买次数 asc;





由此可得出，只购买一次商品的用户购买率=229/671\*100%=34.13%，而购买两次及以上的用户购买率为100%-34.13%=65.87%，因此可推断用户的复购率还是比较高的。

找出购买次数大于等于2的Top10用户

select 用户ID,count(用户ID) as 购买次数

from test1

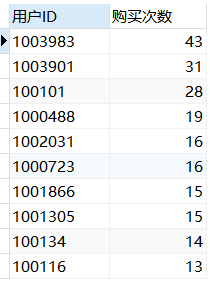
where 行为类型='buy'

group by 用户ID

having count(用户ID)>=2

order by 购买次数 desc

limit 10;



从复购大于等于2的Top10用户，查看这Top10用户购买的Top10商品

select 商品类别ID,count(用户ID) as 购买次数

from test1

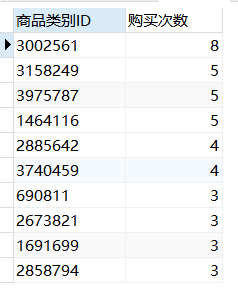
where 用户ID in('1003983','1003901','100101','1000488','1000723','1002031','1001305','1001866','100134','100116')and 行为类型='buy'

group by 商品类别ID

having count(用户ID)>=2

order by 购买次数 desc

limit 10;



由此可知购买次数Top10用户最喜欢购买的Top10商品是商品类目ID为：3002561、1464116等10类商品，这样就可以在推荐商品时，将用户的需求结合起来推荐。

1. **从商品角度**

比较各行为类型的商品

select 商品类别ID,

sum(case when 行为类型 = 'pv' then 1 else 0 end)as 点击量,

sum(case when 行为类型 = 'fav' then 1 else 0 end)as 收藏量,

sum(case when 行为类型 = 'cart' then 1 else 0 end)as 加购量,

sum(case when 行为类型 = 'buy' then 1 else 0 end)as 购买量

from test1

group by 商品类别ID;



创建视图方便查看

create view 商品

as

select 商品类目ID,

sum(case when 行为类型 = 'pv' then 1 else 0 end)as 点击量,

sum(case when 行为类型 = 'fav' then 1 else 0 end)as 收藏量,

sum(case when 行为类型 = 'cart' then 1 else 0 end)as 加购量,

sum(case when 行为类型 = 'buy' then 1 else 0 end)as 购买量

from userbehavior

group by 商品类目ID;

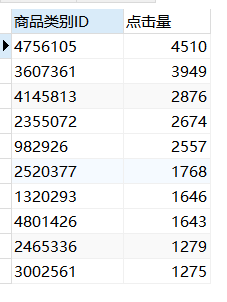
找出点击量-加购量-购买量Top10的商品

select 商品类别ID,点击量

from 商品

order by 点击量 desc

limit 10;



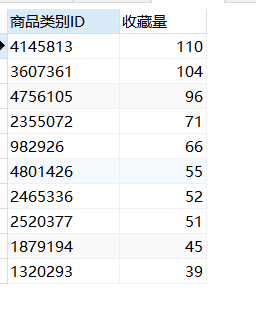
收藏量Top10的商品

select 商品类别ID,收藏量

from 商品

order by 收藏量 desc

limit 10;



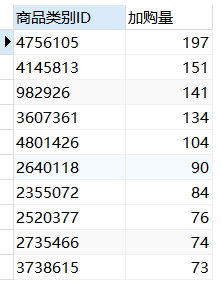
加购量Top10的商品

select 商品类目ID,加购量

from 商品

order by 加购量 desc

limit 10;



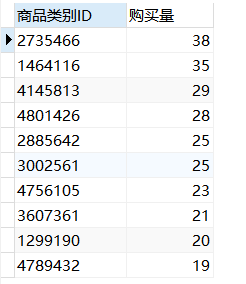
购买量Top10的商品

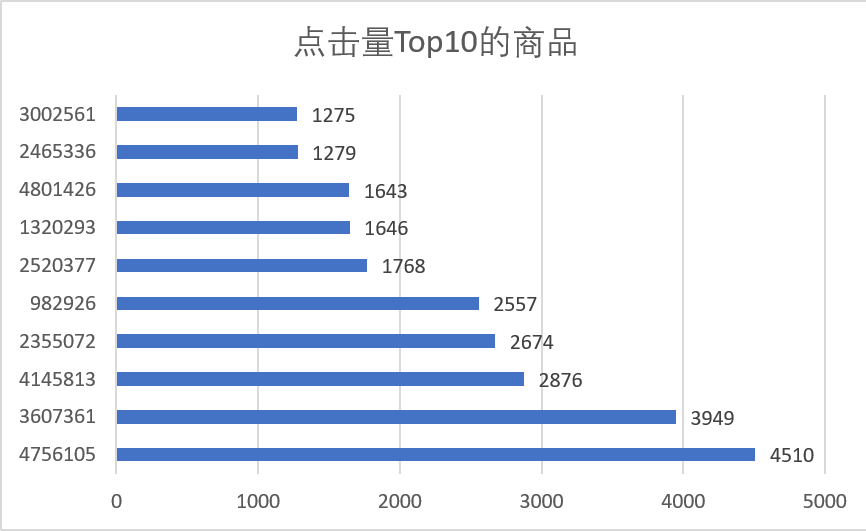
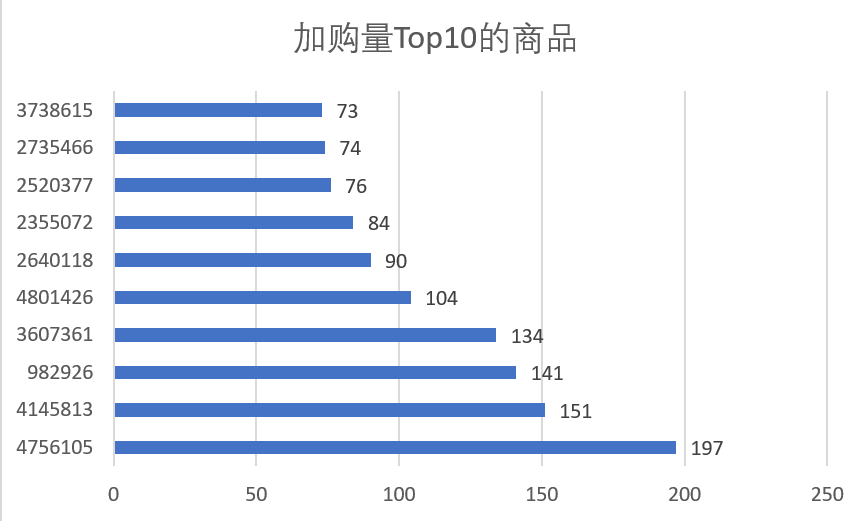
select 商品类别ID,购买量

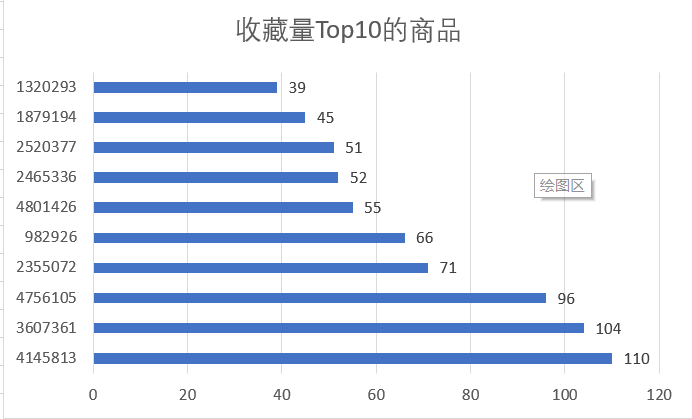
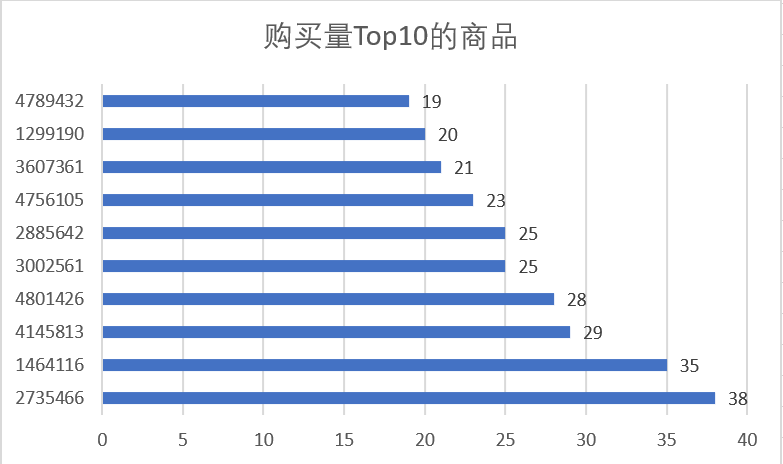
from 商品

order by 购买量 desc

limit 10;



通过对比，我们发现在这Top10商品中，商品类目ID为：4145813、4801426、3607306这3种商品均出现在点击、收藏、加购和购买Top10商品中，这说明点击量Top10商品还有7种商品未被购买，这说明商家在用户的商品喜爱偏好上还没有做到最好的匹配推荐。

通过与复购次数Top10用户最喜欢复购的Top10商品对比发现，点击量Top10、购买量Top10商品与Top10用户最喜欢购买的Top10商品中，仅有商品类目ID：4145813重合。这说明商品的曝光量做得还不够，同时也说明了不同用户对商品的偏好有所不同。

### 类型分析

类型分析即基于RFM分析方法将用户根据其用户价值进行分类，再根据不同类别用户制定不同的营销方案。

1. **计算R、F值**

首先创建视图RFM，存放用户ID，用户最后一次购买的日期与12月3日的间隔，以及用户的购买次数

create view RFM

as

select 用户ID,

datediff('2017-12-03',max(日期)) as '时间间隔R',

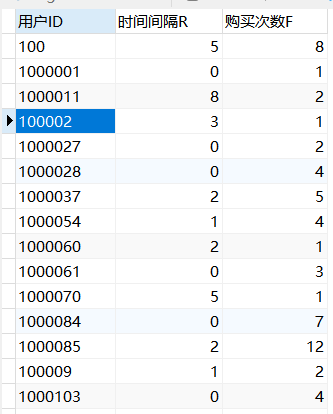
count(行为类型) as '购买次数F'

from test1

where 行为类型='buy'

group by 用户ID

order by 用户ID;



接着从RFM视图中查看最大的时间间隔以及最大的购买次数

select max(时间间隔R),max(购买次数F)

from RFM;



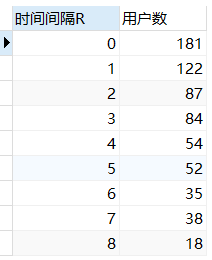
然后从RFM中统计各时间间隔的用户数

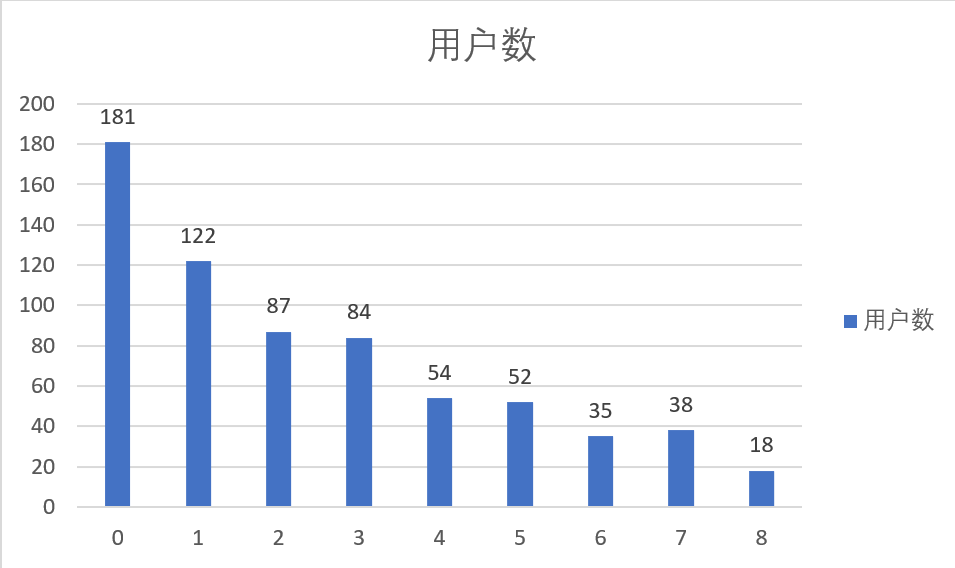
select 时间间隔R,count(用户ID) as 用户数

from RFM

group by 时间间隔R

order by 时间间隔R;





由此可知，购买间隔不超过一天的用户有181人，而购买间隔超过一天的用户数占总用户的绝大部分。

从视图中查询出各购买次数用户数

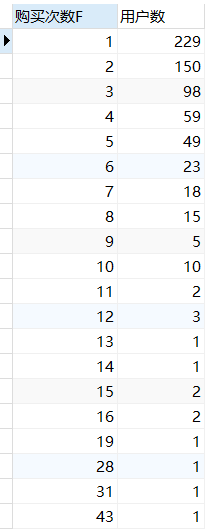
select 购买次数F,

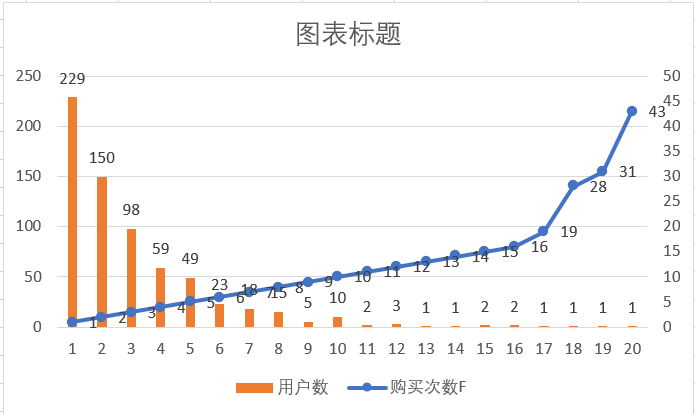
count(用户ID) as 用户数

from RFM

group by 购买次数F

order by 购买次数F;





由图可知，只购买一次的用户数有229人，但购买次数超过一次的用户占绝大部分，用户数目整体趋势随购买次数的增多直线下降成个位数。

1. **给R、F按价值打分**

RFM构建模型的第二步即给R,F按照价值指定打分规则，并创建视图——分数，用于存放R、F值打分。

制定打分规则：构建RFM模型的目的是为了给用户按照其活跃程度进行分类。最近一次消费的时间间隔R越大就表明用户购买时间越远，用户并未经常使用App,分数也就越低;而购买次数F则是用户购物的频率，频率越大则用户越活跃，分数也就越高。

|  |  |  |
| --- | --- | --- |
| **按价值打分** | **最近一次消费时间间隔R** | **购买次数F** |
| 0分 | R<=1 | F<=3次 |
| 1分 |  | 4次<=F<=10次 |
| 2分 | 6<=R<=8 | 11次<=F<=20次 |
| 3分 | 4<=R<=5 | 21次<=F<=30次 |
| 4分 | 2<=R<=3 | 31次<=F<=57次 |

create view 分数 as

select

用户ID,

(case

when 时间间隔R between 0 and 1 then '0分'

when 时间间隔R between 2 and 3 then '4分'

when 时间间隔R between 4 and 5 then '3分'

when 时间间隔R between 6 and 8 then '2分'

else 0

end)

as 'R值打分',

(case

when 购买次数F between 0 and 3 then '0分'

when 购买次数F between 4 and 10 then '1分'

when 购买次数F between 11 and 20 then '2分'

when 购买次数F between 21 and 30 then '3分'

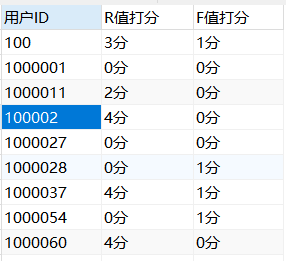
when 购买次数F between 31 and 57 then '4分'

else 0

end)

as 'F值打分'

from RFM;



统计最近一次消费时间间隔的用户数

select count(\*),

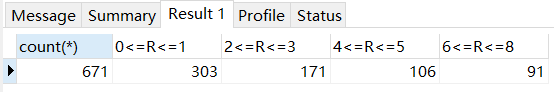
sum(case when R值打分='0分' then 1 else 0 end) as '0<=R<=1',

sum(case when R值打分='4分' then 1 else 0 end) as '2<=R<=3',

sum(case when R值打分='3分' then 1 else 0 end) as '4<=R<=5',

sum(case when R值打分='2分' then 1 else 0 end) as '6<=R<=8'

from 分数;



可以看出来，购买间隔小于等于1的用户占大部分，且随购买间隔越长，用户数逐渐在减少。

统计各购买次数的用户数

select count(\*),

sum(case when F值打分='0分' then 1 else 0 end) as '0<=F=3',

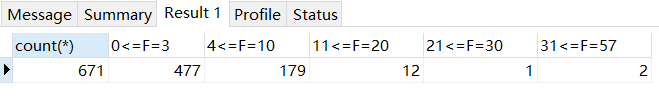
sum(case when F值打分='1分' then 1 else 0 end) as '4<=F=10',

sum(case when F值打分='2分' then 1 else 0 end) as '11<=F=20',

sum(case when F值打分='3分' then 1 else 0 end) as '21<=F=30',

sum(case when F值打分='4分' then 1 else 0 end) as '31<=F=57'

from 分数;

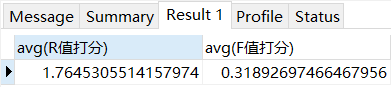


可知，购买次数在3次以内（包括3次）的用户数最多，高达477次，其次是购买次数在4到10次左右的，这说明，大部分用户只购买了几次就不再购买了。

1. **RFM模型第三步：对R、F值打分求平均值**。

select avg(R值打分),avg(F值打分)

from 分数;



**4. RFM模型第四步：用户分类规则**

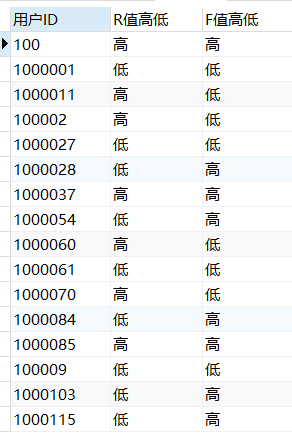
若用户的R值高于R值得平均值则为高，否则为低。F值同理

select 用户ID,

(case when R值打分>(select avg(R值打分)from 分数) then '高' else'低'end) as 'R值高低',

(case when F值打分>(select avg(F值打分)from 分数) then '高' else'低'end) as 'F值高低'

from 分数;



1. **RFM模型第五步：用户分类**

|  |  |  |
| --- | --- | --- |
| R值高低 | F值高低 | 用户分类 |
| 高 | 高 | 价值用户 |
| 高 | 低 | 发展用户 |
| 低 | 高 | 保持用户 |
| 低 | 低 | 挽留用户 |

R值以及F值都高则说明用户近期经常购物，为活跃用户，将其分为价值用户;

R值高说明最近一次消费时间间隔很近，而F值低说明用户购物的频率不高，有发展为价值用户的潜力,将其分为发展用户;

R值低，F值高则说明用户近期没有使用淘宝购物，但是在此之前用户的购物频率很高，需要将其保持稳定下来，将其划分为保持用户;

R值以及FZ值很低的用户则说明该用户近期并不活跃，存在流失的风险，即将其划分为挽留用户。

create view 用户分类划分

as

select 用户ID,

(case when R值打分>(select avg(R值打分)from 分数) then '高' else'低'end) as 'R值高低',

(case when F值打分>(select avg(F值打分)from 分数) then '高' else'低'end) as 'F值高低'

from 分数;

select \*,

( case

when R值高低='高' and F值高低='高' then '价值用户'

when R值高低='高' and F值高低='低' then '发展用户'

when R值高低='低' and F值高低='高' then '保持用户'

when R值高低='低' and F值高低='低' then '挽留用户'

else 0

end

) as'用户分类'

from 用户分类划分;



统计各分类下的用户数

select count(\*) as 总用户数,

sum(case when 用户分类='价值用户' then 1 else 0 end)as 价值用户数,

sum(case when 用户分类='发展用户' then 1 else 0 end)as 发展用户数,

sum(case when 用户分类='保持用户' then 1 else 0 end)as 保持用户数,

sum(case when 用户分类='挽留用户' then 1 else 0 end)as 挽留用户数

from 分类;



用户数按以下大小排序：

发展用户数>挽留用户数>保持用户数>价值用户数，可知价值用户数最少。统计各分类下的用户比例

create view 用户数

as

select count(\*) as 总用户数,

sum(case when 用户分类='价值用户' then 1 else 0 end)as 价值用户数,

sum(case when 用户分类='发展用户' then 1 else 0 end)as 发展用户数,

sum(case when 用户分类='保持用户' then 1 else 0 end)as 保持用户数,

sum(case when 用户分类='挽留用户' then 1 else 0 end)as 挽留用户数

from 分类;

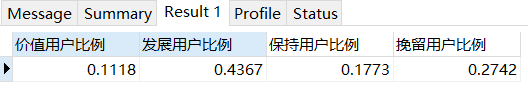
select 价值用户数/总用户数 as 价值用户比例,

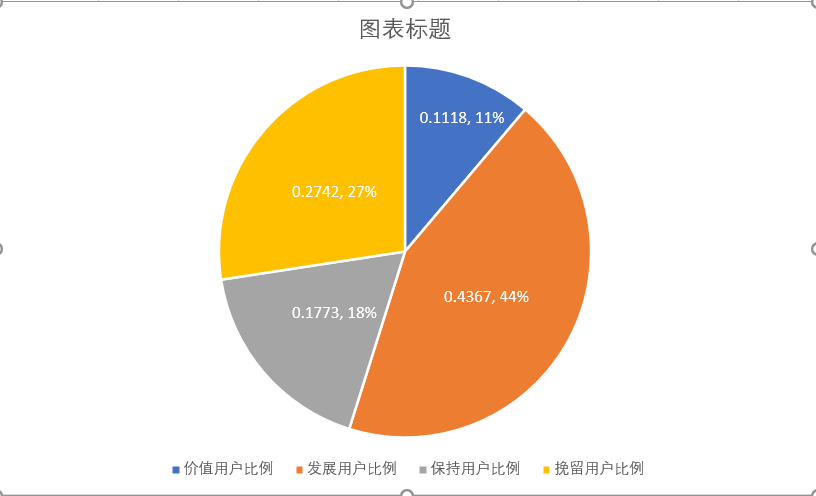
发展用户数/总用户数 as 发展用户比例,

保持用户数/总用户数 as 保持用户比例,

挽留用户数/总用户数 as 挽留用户比例

from 用户数;





综上可知，通过RFM分析方法，我们将数据集中的用户划分为：价值用户、发展用户、保持用户和挽留用户，其中发展用户最多，占总用户数的43.67%，挽留用户、保持用户次之，最后是价值用户，仅占11.18%。

由此，对不同类客户，精细化运营策略如下：

|  |  |
| --- | --- |
| 用户分类 | 精细化运营 |
| 价值用户 | 提高VIP服务，进行满减或打折 |
| 发展用户 | 提高用户的忠诚度，推荐其感兴趣的店铺，刺激消费频率 |
| 保持用户 | 主动联系，赠送其会员等活动优惠 |
| 挽留用户 | 问卷调查，分析原因 |

### 五 结论与建议

1. **获取客户**

邀请朋友点击领取优惠券，或者拼团、砍价之类的，从而增加平台用户数，同时加大商品广告投放，最好选在独立访客量人数集中上线时间，同步开展营销活动，关注新增客户指标，降低获客成本，以此吸引更多用户进入平台。

1. **激活用户**

当平台中新进入的用户数多了，但是使用率却很低，这时候就需要去激活用户。而激活用户需要关注产品的“啊哈时刻”和各业务流程用户的流失率，通过分析我们发现用户流失主要发生在点击-加购或收藏这一环节。我们猜测可能是因为用户点击加入购物车或收藏的过程太过于复杂，导致用户不愿继续进行下一步操作导致的，这时我们可以对用户加入购物车或收藏的过程提出一些优化建议：

1. 首页直接推荐商品，不放无效信息，不需要用户点击多次才到商品页面；
2. 不设置购物车，点击商品后直接购买支付，减少用户犹豫时间；
3. 先付款后拼团，将支付流程提前，系统自动将所有用户以团购价自动拼团；
4. 通过游戏娱乐，给用户发放奖励来唤醒用户，例如：砍一刀减价等。

**3. 提高留存**

我们知道要想提高留存就是要培养用户的习惯，这里可采取的措施有：

1. 对所有商品免邮，让用户购物中习惯免邮费，给购买次数多且金额大的用户打折，让用户习惯会员折扣，从而不再去别的平台购买东西。
2. 设置新用户奖励和一定复购次数用户折扣，让用户能从获取平台优惠券或者积分，减少付款金额。

**4. 增加收入**

将收入分为服务收入和广告收入，找损失潜在收益的地方，分析用户关键环节放弃的原因，细化解决相应问题。

在服务收入上，根据RFM分析方法确定的用户类型，对不同的用户采取不同的措施，从而提高复购率、成交量：

1. 价值用户:提高VIP服务，当这类用户再次在平台购买时，给予一定的VIP折扣：满减或者打折；
2. 发展用户:其购买次数低，这时候要想办法提高其购买频率：通过短信方式，适当地赠送优惠券；
3. 保持用户:这类用户购买次数高，但是购买时间间隔太长，属于一段时间没来的忠实用户，这时候应该主动与用户保持联系，提高复购率；
4. 挽留用户:这类用户购买次数少且时间间隔长，属于即将流失的用户，这时候需要主动联系用户，弄清楚原因，想办法挽留。

在广告收入上，根据之前分析的Top10部分，找出用户偏好，投放需求度高的商品广告和营销活动，尽可能放在流量多的时间段进行推荐和宣传。

**5. 推荐**

针对淘宝平台，让用户推荐给其他人的方案有：

1）深入了解用户痛点，梳理平台配送地点和时间，优化用户选购流程，加快收货时间

2）推出拉新活动，老用户分享介绍新用户注册，给予随机赠品奖励

3）好物推荐，用户分享购物链接到微信、朋友圈等平台，均可获取优惠券