淘宝用户行为分析案例——天池数据集 User Behavior Data from Taobao

说明

本次数据分析基于阿里云天池数据集（用户行为数据集），使用转化漏斗，RFM 模型，对常见电商分析指标，包括转化率，PV,UV,留存率等进行分析，分析过程中使用 MySQL 以及 Navicat 进行数据预览与清洗处理，使用 Excel 进行数据可视化。

一、数据集与分析目的

1、数据集

字段	说明
user_id	整数类型，序列化后的用户ID
item_id	整数类型，序列化后的商品ID
category_id	整数类型，序列化后的商品所属类目ID
behaviortype	字符串，枚举类型，包括('pv', 'buy', 'cart', 'fav')
timestamps	行为发生的时间戳

该数据集记录用户在淘宝网站浏览商品产生的行为信息。

2、分析目的

1）了解网站流量情况；

2）了解该阶段用户粘性以及用户行为习惯；

3）了解商品销售情况；

分析框架如下：

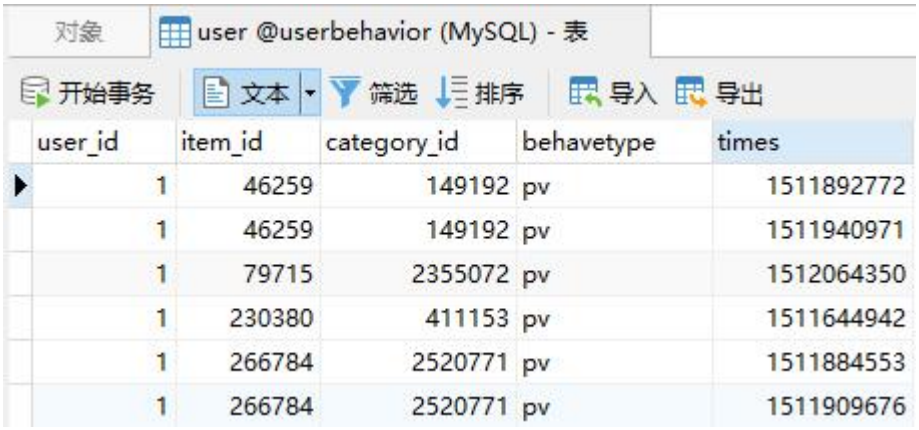
二、数据处理

1、数据预览

MySQL 建表，列名重命名：

```
mysql> use userbehavior;
Database changed
mysql> create table user(
    -> user_id int not null,
    -> item_id int not null,
    -> category_id int not null,
    -> behavetype varchar(10) not null,
    -> times int not null,
    -> constraint user_behave primary key(user_id,item_id,times));
Query OK, 0 rows affected (2.49 sec)
```

使用 Navicat 导入数据：



user_id	item_id	category_id	behavetype	times
1	46259	149192	pv	1511892772
1	46259	149192	pv	1511940971
1	79715	2355072	pv	1512064350
1	230380	411153	pv	1511644942
1	266784	2520771	pv	1511884553
1	266784	2520771	pv	1511909676

2、数据处理

（1）日期处理

将行为发生时间转为 datetime 类型，获取其发生日期及时间：

```
SET SQL_SAFE_UPDATES = 0;
ALTER TABLE user ADD COLUMN datentime TIMESTAMP(0) NULL;
UPDATE user
SET datentime = FROM_UNIXTIME(times);
ALTER TABLE user ADD COLUMN dates CHAR(10) NULL;
UPDATE user
SET dates = SUBSTRING(datentime FROM 1 FOR 10);
ALTER TABLE user ADD COLUMN hours CHAR(10) NULL;
UPDATE user
SET hours = SUBSTRING(datentime FROM 12 FOR 2);
```

（2）选取数据：本次分析针对 2017-11-25 至 2017-12-03 数据，对其余数据进行删除处理，共删除 528920 行

```
DELETE from user
where datentime<'2017-11-25 00:00:00'
or datentime>'2017-12-04 00:00:00';
```

（3）查看缺失值

```
SELECT * from user
where user_id is null
or item_id is null
or category_id is null
or behavetype is null
or times is null;
```

信息 | 结果 1 | 剖析 | 状态

user_id	item_id	category_id	behavetype	times	datentime	dates
(N/A)	(N/A)	(N/A)	(N/A)	(N/A)	(N/A)	(N/A)

数据集中不存在缺失值。

（4）查看异常值：主要查看是否存在异常时间

```
SELECT max(datentime),min(datentime),max(dates),min(dates)
from user;
```

信息 | 结果 1 | 剖析 | 状态

max(datentime)	min(datentime)	max(dates)	min(dates)
2017-12-03 23:59:59	2017-11-25 00:00:0(	2017-12-03	2017-11-25

日期不存在异常值。

三、数据分析

1、基于转化漏斗模型分析用户行为

创建用户行为视图：

```
create view userbehave as
select user_id,count(behavetype) num,sum(if(behavetype='pv',1,0)) pv,
sum(if(behavetype='cart',1,0)) cartbox,sum(if(behavetype='fav',1,0))
favor,sum(if(behavetype='buy',1,0)) buy,dates,hours
from user
group by user_id,dates,hours
order by user_id,dates,hours;
```

user_id	num	pv	cartbox	favor	buy	dates	hours
1	1	1	0	0	0	2017-11-25	06
1	1	1	0	0	0	2017-11-25	09
1	2	2	0	0	0	2017-11-25	15
1	1	1	0	0	0	2017-11-25	21

（1）常见电商指标分析

1）获客：每日新增用户数

定义首次购买的用户为新增客户：

```
SELECT count(DISTINCT t.user_id) as '新增用户数',t.日期 from
(SELECT user_id,min(dates)as '日期' from user
WHERE behavetype='buy' GROUP BY user_id)t
GROUP BY t.日期;
```

信息 | 结果 1 | 剖析 | 状态

新增用户数	日期
4965	2017-11-25
3929	2017-11-26
3475	2017-11-27
2793	2017-11-28
2447	2017-11-29
2245	2017-11-30
1849	2017-12-01
2029	2017-12-02

从 2017-11-25 日开始，每天都有新增用户，11 月 25 日的新增用户最多。

2）激活：用户数、商品数、类目数

```
SELECT count(DISTINCT user_id) as num_user,
COUNT(DISTINCT item_id) as num_item,
COUNT(DISTINCT category_id) as num_category
from user;
```

信息 | 结果 1 | 剖析 | 状态

num_user	num_item	num_category
37376	854126	6966

从整体上看，本次分析共涉及 37376 名用户，近 86 万商品，近 7000 个品类。

3）PV、UV、日均 PV

总体访问量 2958330 人次，独立访客数（UV）37376 人次

```
33  SELECT count(*) as 总体访问量 from user
34  WHERE behavetype='pv';
```

信息 | 结果 1 | 剖析 | 状态

总体访问量
2958330

日均访问量：12 月 2 日日均访问量最大，其余时间日均访问量较稳定，均维持在 35 万-38 万人次左右

```
35  SELECT dates,count(user_id) as '日均访问人数'
36  from user
37  WHERE behavetype='pv'
38  GROUP BY dates
39  ORDER BY dates;
```

信息 | 结果 1 | 剖析 | 状态

dates	日均访问人数
2017-11-25	354563
2017-11-26	364526
2017-11-27	341416
2017-11-28	338412
2017-11-29	349887
2017-11-30	358866
2017-12-01	373806
2017-12-02	476848

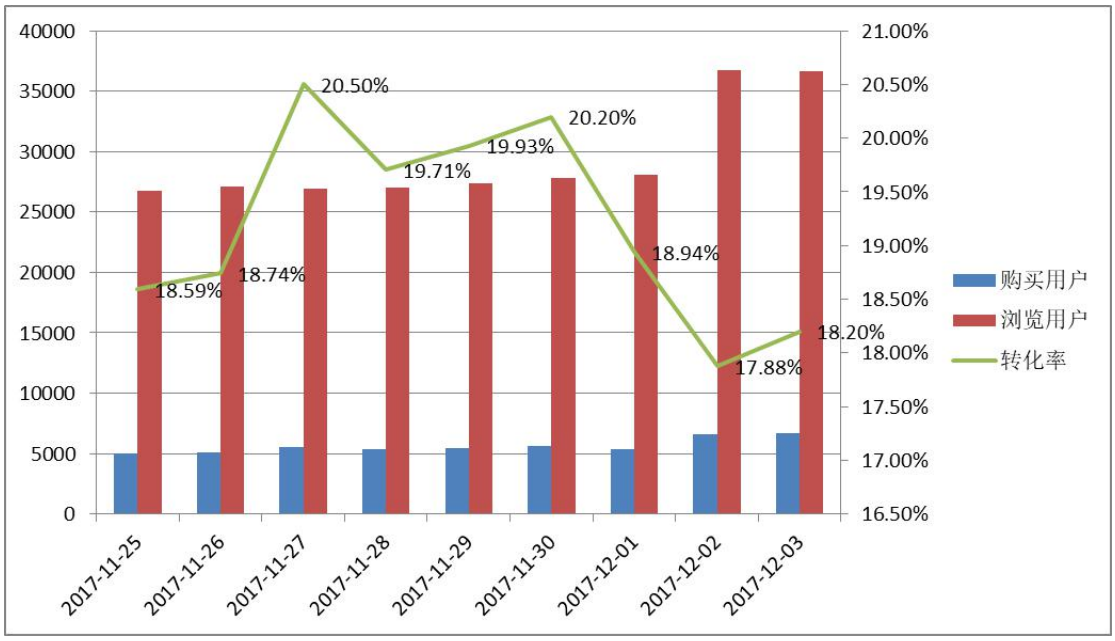
4）购买转化率：支付访客数/总访客数

购买转化率在 27 日-30 日之间较高，30 日后转化率迅速下跌：

```
36  select a.dates,a.ac,b.au,concat(round(a.ac/b.au*100,2),'%')
    conversion
37  from
38  (SELECT dates,count(distinct user_id) ac FROM userbehavior.
    userbehave where buy<>0 group by dates) as a
39  join
40  (SELECT dates,count(distinct user_id) au FROM userbehavior.
    userbehave group by dates) as b
41  on a.dates=b.dates;
```

信息 | 结果 1 | 剖析 | 状态

dates	ac	au	conversion
2017-11-2	4965	26710	18.59%
2017-11-2	5080	27107	18.74%
2017-11-2	5512	26892	20.50%



5）留存：次日留存回购人数、3 日留存回购人数

次日留存回购人数：

```
45  SELECT t1.dates,COUNT(t1.dates),COUNT(t2.dates),concat(round(COUNT(
    t2.dates)/count(t1.dates)*100,2),'%') as 留存率 from
46 (SELECT user_id,dates from user where behavetype='buy' group by
47  user_id,dates)t1 LEFT JOIN
48 (SELECT user_id,dates from user where behavetype='buy' group by
49  user_id,dates)t2
50  on t1.user_id=t2.user_id and t1.dates=DATE_SUB(t2.dates,INTERVAL 1
    day) GROUP BY t1.dates;
```

信息 | 结果 1 | 剖析 | 状态

dates	COUNT(t1.dates)	COUNT(t2.dates)	留存率
2017-11-27	5512	1318	23.91%
2017-11-25	4965	1151	23.18%
2017-11-28	5318	1230	23.13%
2017-11-26	5080	1260	24.80%
2017-11-29	5459	1319	24.16%
2017-12-02	6563	1700	25.90%
2017-12-01	5316	1395	26.24%
2017-11-30	5624	1243	22.10%
2017-12-03	6665	0	0.00%

当天购买后，3 天后再次购买的人数：

```
53  SELECT t1.dates,COUNT(t1.dates),COUNT(t2.dates),concat(round(COUNT(
    t2.dates)/count(t1.dates)*100,2),'%') as 留存率 from
54 (SELECT user_id,dates from user where behavetype='buy' group by
55  user_id,dates)t1 LEFT JOIN
56 (SELECT user_id,dates from user where behavetype='buy' group by
57  user_id,dates)t2
58  on t1.user_id=t2.user_id and t1.dates=DATE_SUB(t2.dates,INTERVAL 2
    day) GROUP BY t1.dates;
```

信息 | 结果 1 | 剖析 | 状态

dates	COUNT(t1.dates)	COUNT(t2.dates)	留存率
2017-11-27	5512	1216	22.06%
2017-11-25	4965	1135	22.86%
2017-11-28	5318	1182	22.23%
2017-11-26	5080	1120	22.05%
2017-11-29	5459	1122	20.55%
2017-12-02	6563	0	0.00%
2017-12-01	5316	1355	25.49%
2017-11-30	5624	1352	24.04%
2017-12-03	6665	0	0.00%

整体上看，次日留存率在 22%-26%之间，3 日留存率在 22%-25%之间，说明在数据集时间范围内回购率还是很高的。

6）复购：用户复购次数

```
SELECT count(t.购买次数),COUNT(if(t.购买次数>1,1,null)),
COUNT(if(t.购买次数>1,1,null))/count(t.购买次数) as '复购率'
from
(SELECT user_id,count(user_id) as '购买次数' from user
where behavetype='buy' GROUP BY user_id)t;
```

信息 | 结果 1 | 剖析 | 状态

count(t.购买次数)	COUNT(if(t.购买次数>1,1,r	复购率
23732	14779	0.6227

超过 62%的用户购买了一次以上。

7）跳失率：只有浏览行为的用户数/总用户数

```
SELECT SUM(if(t.fav_num=0 and t.cart_num=0 and t.buy_num=0,1,0)) as
 '只有浏览行为的用户数',count(t.user_id) as '总用户数',CONCAT(ROUND
(SUM(if(t.fav_num=0 and t.cart_num=0 and t.buy_num=0,1,0))/count(t.
user_id)*100,2),'%')as '跳失率' from
(SELECT user_id,sum(if(behavetype='fav',1,0))as fav_num,sum(if(
behavetype='cart',1,0)) as cart_num,sum(if(behavetype='buy',1,0))
as buy_num from user GROUP BY user_id) t;
```

信息 | 结果 1 | 剖析 | 状态

只有浏览行为的用户数	总用户数	跳失率
2197	37376	5.88%

跳失率仅有 5.88%，说明页面对用户的吸引力较强。

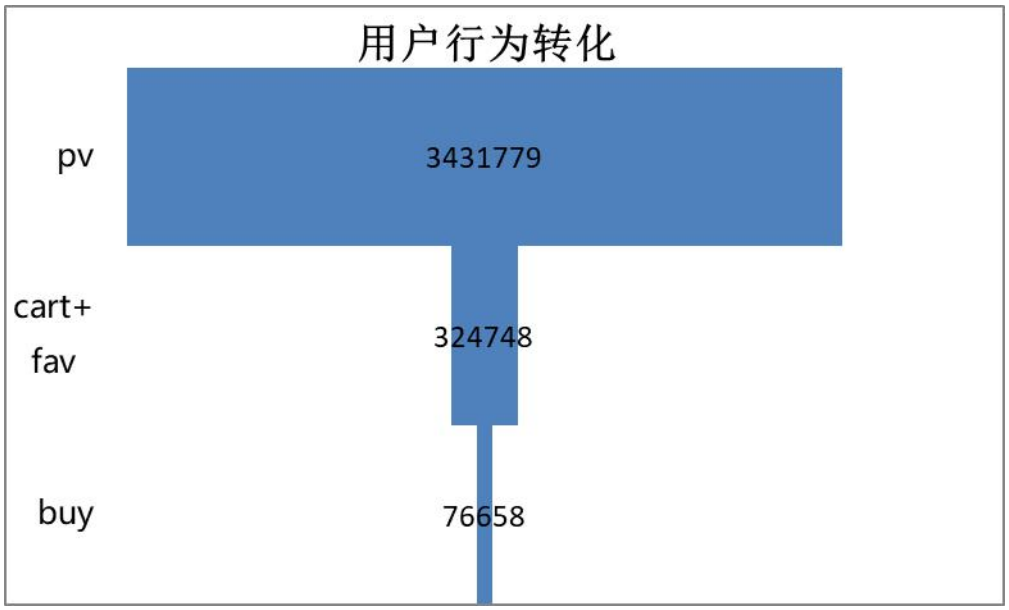
（2）用户行为转化漏斗模型分析

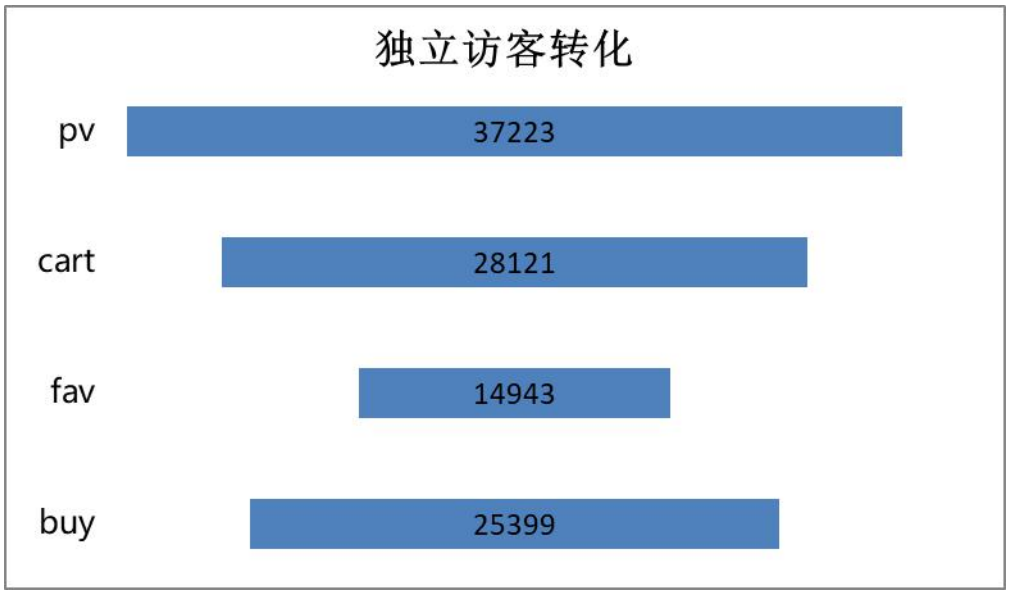
```
SELECT behavetype as 阶段,count(1) as 人数,count(distinct user_id)
as 用户ID FROM user
group by behavetype
order by count(1) desc;
```

信息 | 结果 1 | 剖析 | 状态

阶段	人数	用户ID
pv	3431779	37223
cart	213608	28121
fav	111140	14943
buy	76658	25399

用户行为转化中，将收藏和加入购物车的行为整合为第二阶段：





独立访客转化率达到了 68%，但从用户行为来看，浏览到购买的转化率只有 2.23%。

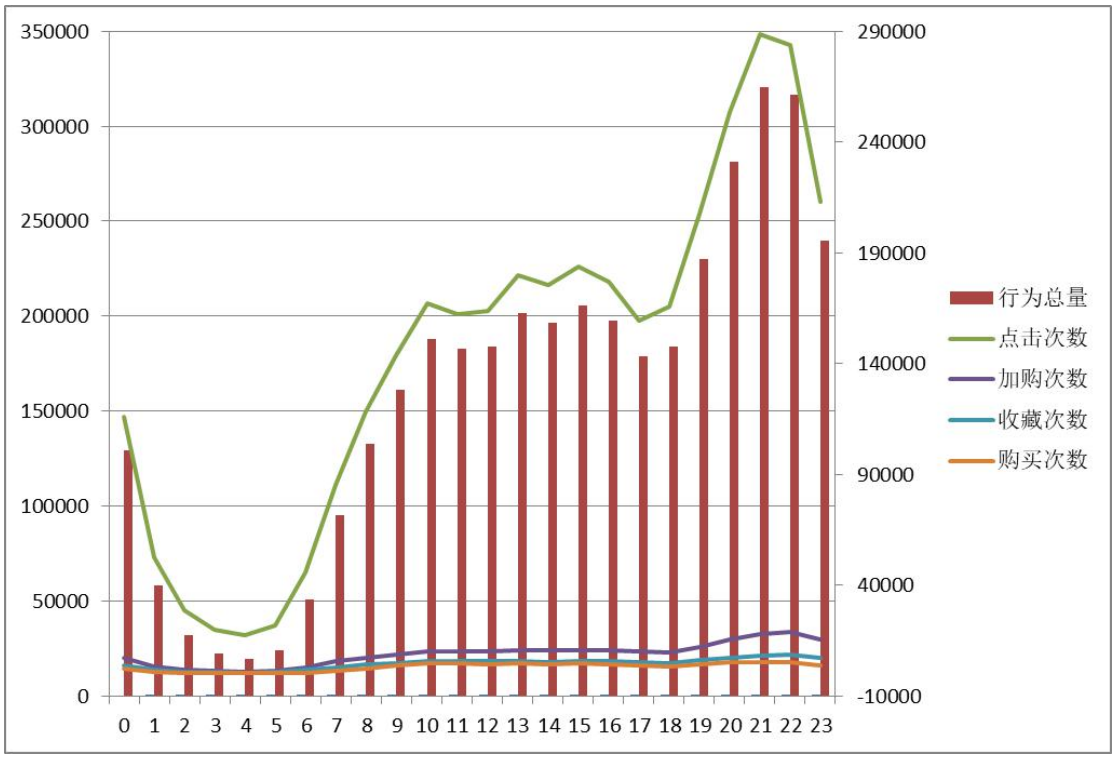
2、从时间维度分析用户行为

一天中用户活跃时段分布：

```
59  SELECT a.hours,count(behavetype),
60  (SELECT count(*) from user where hours=a.hours and behavetype='pv')
     as '点击次数', (SELECT count(*) from user where hours=a.hours and
    behavetype='cart') as '加购次数',(SELECT count(*) from user where
    hours=a.hours and behavetype='fav') as '收藏次数',(SELECT count(*)
    from user where hours=a.hours and behavetype='buy') as '购买次数'
61  from user a
62  GROUP BY hours
63  order by hours;
```

信息 | 结果 1 | 剖析 | 状态

hours	count(behavetype)	点击次数	加购次数	收藏次数	购买次数
00	129223	115985	7058	3949	2231
01	58415	52596	3251	1720	848
02	32116	28850	1729	1055	482
03	22300	20048	1310	706	236



可以看出：

1）每日 0 点到 4 点，用户活跃度快速降低，降到一天中的最低值，5 点到 10 点用户活跃度逐渐上升；

2）下午整体时段的活跃度较平稳，4 点到 6 点有所回落；

3）用户整体在晚上活跃度高，在 9 点-10 点达到一天中的最高峰，活跃度是上午的 2 倍左右。

3、从商品维度分析用户行为

1）热销商品、热销类别

查看销量 TOP20 的商品，其中销量前三的商品为 3122135、3031354、1910706：

```
303  SELECT item_id,count(DISTINCT user_id) from user
304  where behavetype='buy'
305  GROUP BY item_id
306  ORDER BY count(DISTINCT user_id) desc
307  LIMIT 20;
```

信息 | 结果 1 | 剖析 | 状态

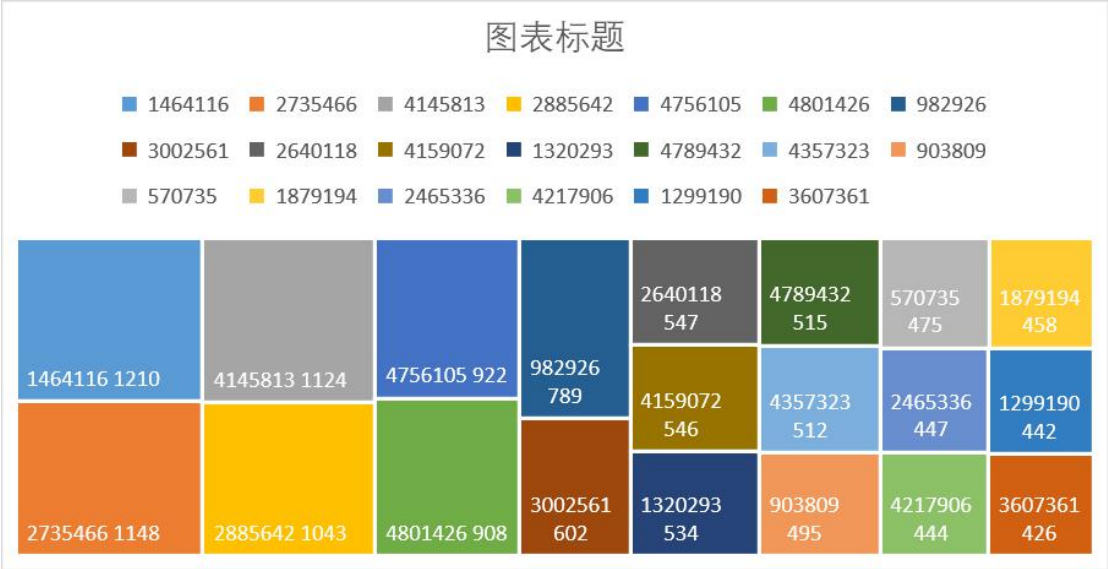
item_id	count(DISTINCT user_id)
3122135	58
3031354	30

从类别上看，1464116 标签号类别销量最好，共有 1210 个用户购买该类别商品：

```sql
SELECT category_id,COUNT(DISTINCT user_id) from user
where behavetype='buy'
GROUP BY category_id
order by COUNT(DISTINCT user_id) desc
limit 20;
```

信息 | 结果 1 | 剖析 | 状态

category_id	COUNT(DISTINCT user_id
1464116	1210
2735466	1148
4145813	1124

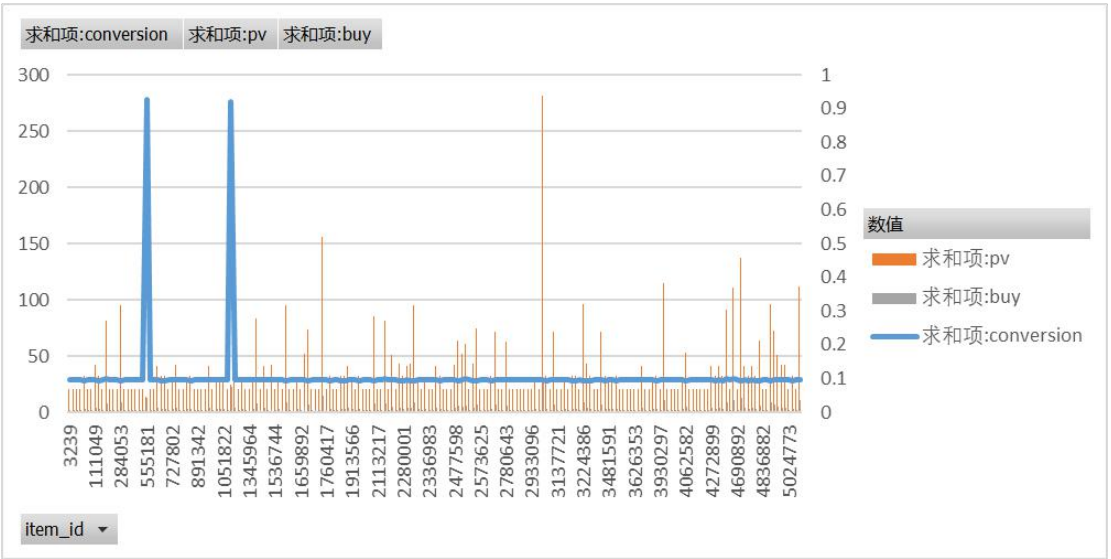


2）商品购买转化率

```sql
CREATE VIEW onitem as
SELECT item_id,category_id,sum(if(behavetype='pv',1,0)) pv,sum(if(
behavetype='buy',1,0)) buy from user GROUP BY item_id,category_id;
SELECT item_id,category_id,pv,buy,concat(round(buy/pv*100,2),'%')
conversion
FROM onitem
order by conversion  desc
limit 200;
```

信息 | 结果 1 | 剖析 | 状态

item_id	category_id	pv	buy	conversion
555181	194104	14	13	92.86%
1116492	2297500	25	23	92.00%
4681909	570735	111	11	9.91%
4554568	3158249	91	9	9.89%



商品编号为 555181 和 1116492 的商品购买转化率达到了 92%。

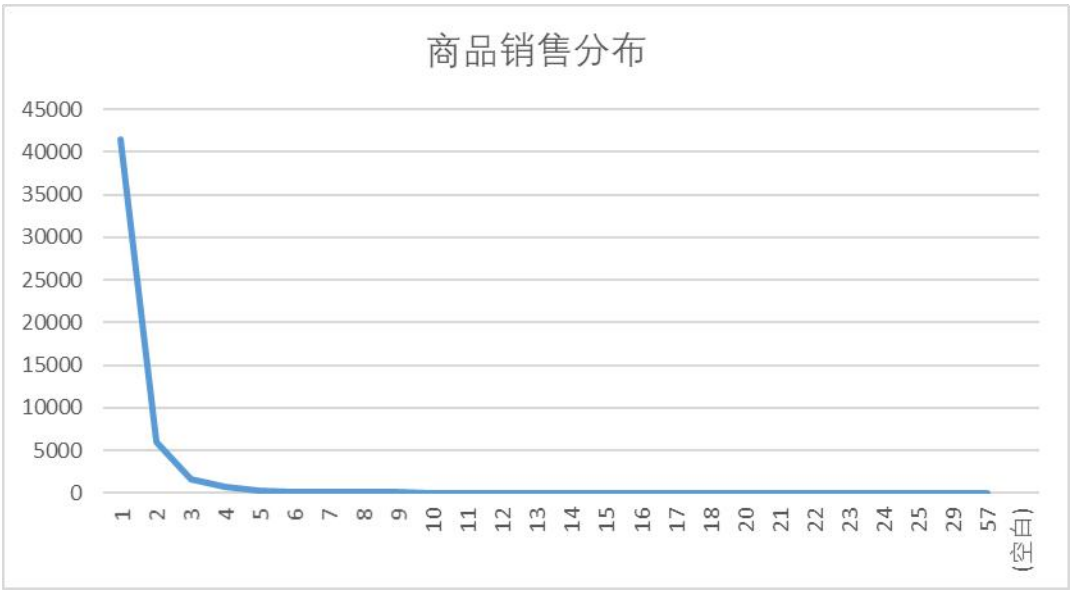
3）商品销售分布

按照商品销量对商品分类统计：

```
45  SELECT a.购买次数,count(item_id) as '商品量'
46  from
47 (SELECT item_id,count(behavetype) as '购买次数'
48  from user
49  where behavetype='buy'
50  GROUP BY item_id
51  ORDER BY 购买次数) a
52  GROUP BY a.购买次数
53  ORDER BY 商品量 desc;
```

信息 | 结果 1 | 剖析 | 状态

购买次数	商品量
1	41410
2	5963
3	1603



大部分销售来源于销售次数为 1 的 41410 个商品，说明该时段销售主要源于长尾分布，而非销售数量较高的畅销品。

4、基于 RFM 模型分析用户行为

借鉴 RFM 模型，暂时不考虑 M 维度（由于数据集中没有给出具体金额），对 R（最近一次购买时间）F（购买频率）进行分析，完成用户分层。

1）最近一次消费时间和消费频率

```
66  SELECT user_id,DATEDIFF('2017-12-05',max(dates)) as
    '最后一次交易距今时间',count(behavetype) as '交易次数'
67  from user
68  where behavetype='buy'
69  GROUP BY user_id;
```

信息 | 结果 1 | 剖析 | 状态

user_id	最后一次交易距今时间	交易次数
100	7	8
117	7	10
119	6	3

最后一次交易距今时间和交易次数的最大值：

```
70  SELECT max(t.最后一次交易距今时间),max(t.交易次数) from
71 (SELECT user_id,DATEDIFF('2017-12-05',max(dates)) as
    '最后一次交易距今时间',count(behavetype) as '交易次数'
72  from user
73  where behavetype='buy'
74  GROUP BY user_id) t;
```

信息 | 结果 1 | 剖析 | 状态

max(t.最后一次交易距今时	max(t.交易次数)
10	84

2）对用户进行评分

```
230 SELECT user_id,(case WHEN t.最后一次交易距今时间 in (0,1) THEN 4
231                     when t.最后一次交易距今时间 in (2,3) THEN 3
232                     WHEN t.最后一次交易距今时间 in (4,5) THEN 2
233                     WHEN t.最后一次交易距今时间 in (6,7) THEN 1
234                     WHEN t.最后一次交易距今时间>7 THEN 0 ELSE null
    end) as Recent,(case when t.交易次数>=16 then 4
235                      when t.交易次数 between 11 and 15 then 3
236                      when t.交易次数 BETWEEN 6 and 10 then 2
237                      when t.交易次数 BETWEEN 1 and 5 then 1
```

信息 | 结果 1 | 剖析 | 状态

user_id	Recent	Frequent
220615	3	1
220626	3	1
220628	3	1

导出计算得 R 的平均值为 1.99，F 的平均值为 1.15。用均值来划分 4 个客户层次：

F（3-4）	重要保持客户	重要价值客户
F（1-2）	重要挽留客户	重要发展客户
	R（1-2）	R（3-4）

3）用户分层

```
SELECT user_id,Recent,Frequent,
(CASE WHEN Frequent <= 1.15 AND Recent <= 1.99 THEN '重要挽留客户'
WHEN Frequent <= 1.15 AND Recent > 1.99 THEN '重要发展客户' WHEN
Frequent > 1.15 AND Recent <= 1.99 THEN '重要保持客户' WHEN
Frequent > 1.15 AND Recent > 1.99 THEN '重要价值客户' END ) AS
'客户分类' FROM rmf ORDER By Recent DESC,Frequent DESC;
```

信息 | 结果 1 | 剖析 | 状态

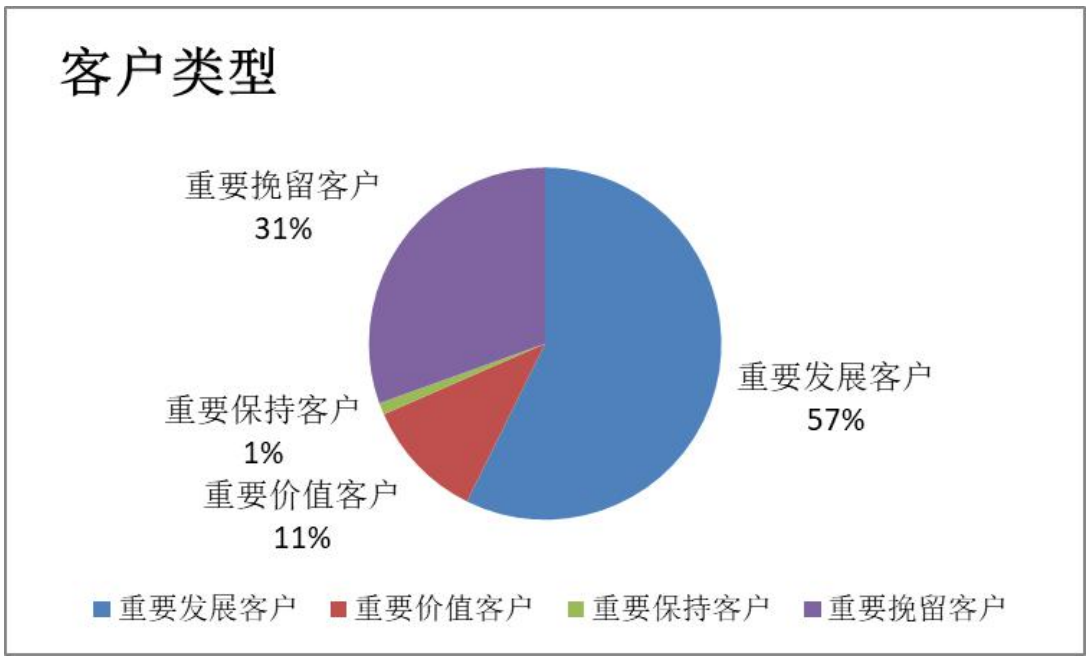
user_id	Recent	Frequent	客户分类
171362	3	4	重要价值客户
172091	3	4	重要价值客户
173487	3	4	重要价值客户

各类客户数量：

```
SELECT t.客户分类,COUNT(t.user_id) from
(SELECT user_id,Recent,Frequent,
(CASE WHEN Frequent <= 1.15 AND Recent <= 1.99 THEN '重要挽留客户'
WHEN Frequent <= 1.15 AND Recent > 1.99 THEN '重要发展客户' WHEN
Frequent > 1.15 AND Recent <= 1.99 THEN '重要保持客户' WHEN
Frequent > 1.15 AND Recent > 1.99 THEN '重要价值客户' END ) AS
'客户分类' FROM rmf ORDER By Recent DESC,Frequent DESC)t
GROUP BY t.客户分类;
```

信息 | 结果 1 | 剖析 | 状态

客户分类	COUNT(t.user_id)
重要发展客户	14558
重要价值客户	2820
重要保持客户	277
重要挽留客户	7744



四、结论

1、基于转化漏斗模型分析用户行为：

- 2017-11-25 至 2017-12-03 期间总计近 300 万流量，86 万商品，近 7000 个品类，37376 的独立访客，日均访问量在 35 万-38 万之间；
- 数据集期间内新增用户逐渐减少，购买转化率在 27 日-30 日之间较高，30 日后转化率迅速下跌，可能与即将到来的双 12 大促有关；
- 从用户行为来看，浏览到购买的转化率只有 2.23%，说明用户花费大量时间在产品搜索上，因此考虑优化搜索信息，减少搜索对比频率，提高行为转化率；
- 独立访客的购买转化率较高，达到了 68%，考虑增加产品曝光，获取并激活更多用户；
- 次日回购及 3 日回购均保持较高水平，用户复购率达到 62%以上，说明用户黏性较高。对于流失用户，通过分析流失用户特征，结合用户画像，通过各种活动进行用户唤醒，提高用户留存。

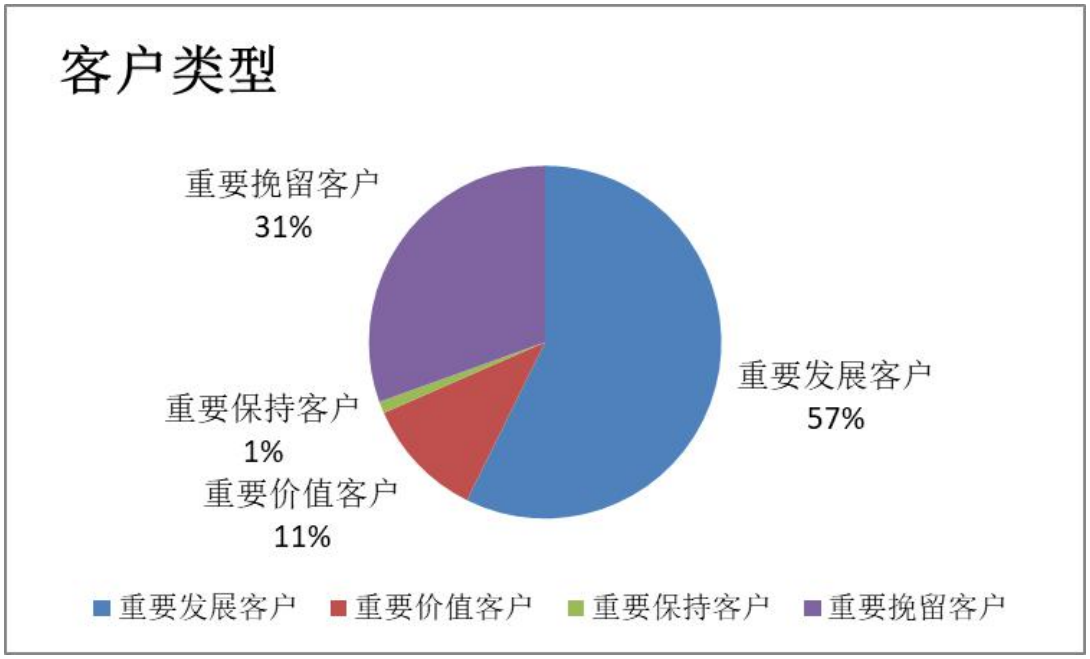
2、从时间维度分析用户行为：

每天晚上 9 点-11 点是用户活跃的高峰期，在制定运营策略时，可以利用这个规律来进行创收，选择在该时间段推出各种网店直播等互动营销手段。

3、从商品维度分析用户行为：

- 3122135、3031354、1910706 等商品销量较高，从类别上看，1464116 标签号类别销量最好，可以着重推广这几种产品；
- 淘宝平台商品售卖主要是依靠长尾商品的累计效应，并非爆款商品的带动。商家其实也可以通过增大宣传力度、突出产品优势等方式来打造爆款商品并获利。

4、基于 RFM 模型分析用户行为：



- 重要发展客户所占比重最大，运营活动可以重点针对这部分用户，通过积分兑换、拼团打折等活动引起用户注意；
- 重要挽留客户：可以通过推送提醒，短信召回等含促销活动的字眼进行吸引唤回；
- 重要价值客户：需要重点关注，活动投放时需谨慎对待，不要引起用户反感；
- 重要保持客户：制定相应的运营策略来保持用户粘性。