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，AARRR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。由于数据集过大，选取其中105万条数据进行分析。

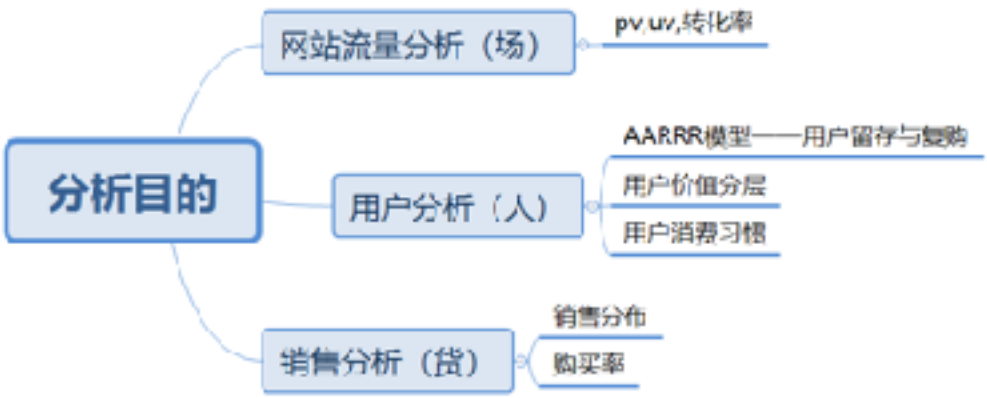
2.分析目的

1.了解网站流量情况；

2.了解该阶段网站用户粘性以及用户行为习惯；

3.了解网站商品销售情况；

分析逻辑如下：



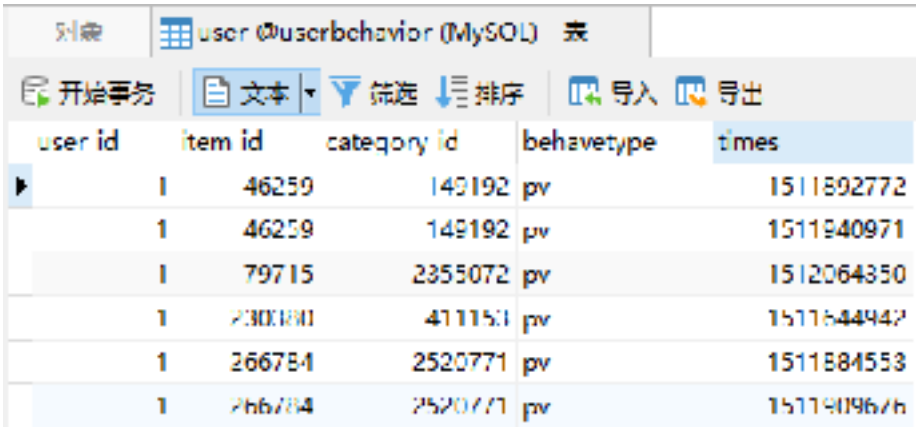
二.数据处理

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 (2.69 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	200000	411153	pv	1511644442
1	266784	2520771	pv	1511884553
1	266784	2520771	pv	1511940962

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

2.选取数据：选取2017-11-25至2017-12-04之间的数据进行分析

```
DELETE from user
where datetime<'2017-11-25 00:00:00'
or datetime>'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	datetime	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(datetime),min(datetime),max(dates),min(dates)
from user;
```

信息 结果 1 剖析 状态

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

日期不存在异常值。

三.数据提取与分析

1.网站流量分析

创建用户行为视图：

```
create view userbenave 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.用户参与度（PV）分析

```
SELECT DISTINCT user_id,dates
from user
where behavetype='pv'
ORDER BY dates;
```

信息 结果 1 剖析 状态

user_id	dates
1	2017-11-2
19	2017-11-2

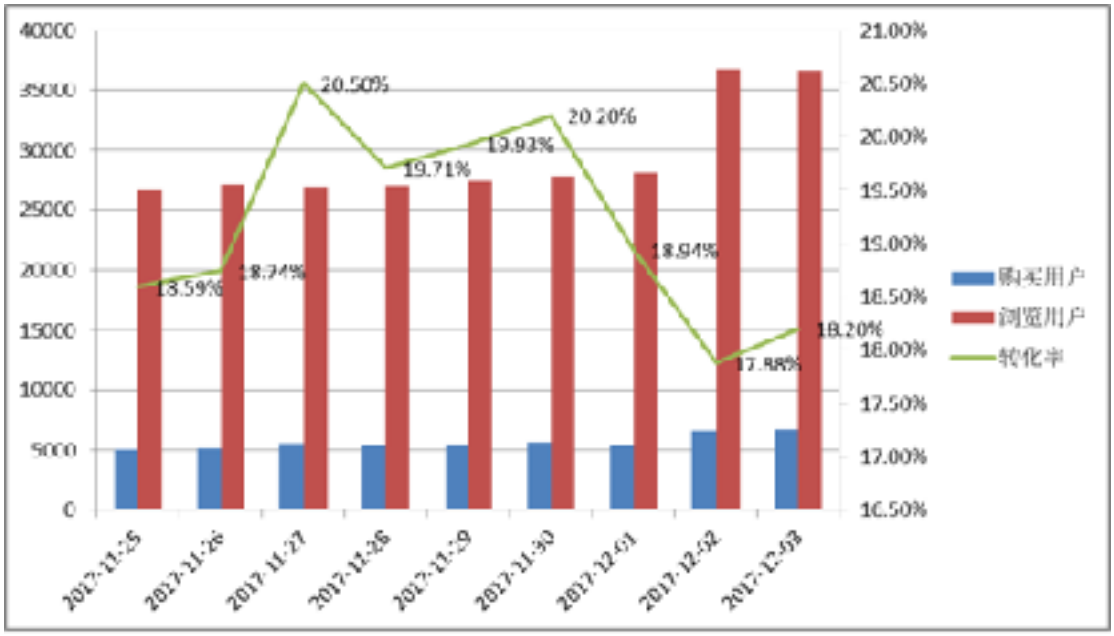
分析：11月25日至12月1日用户数变化幅度较小，范围在25000-30000之间，从12月2日开始有较大幅度增长，对比11月25，26（周末）有较大提升，可能与周末的双十二营销活动有关。

2.网站购买转化率（UV）：（支付转化率=支付访客数/总访客数）

```
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%



分析：支付转化率在27日-30日之间较高，30日后转化率迅速下跌。

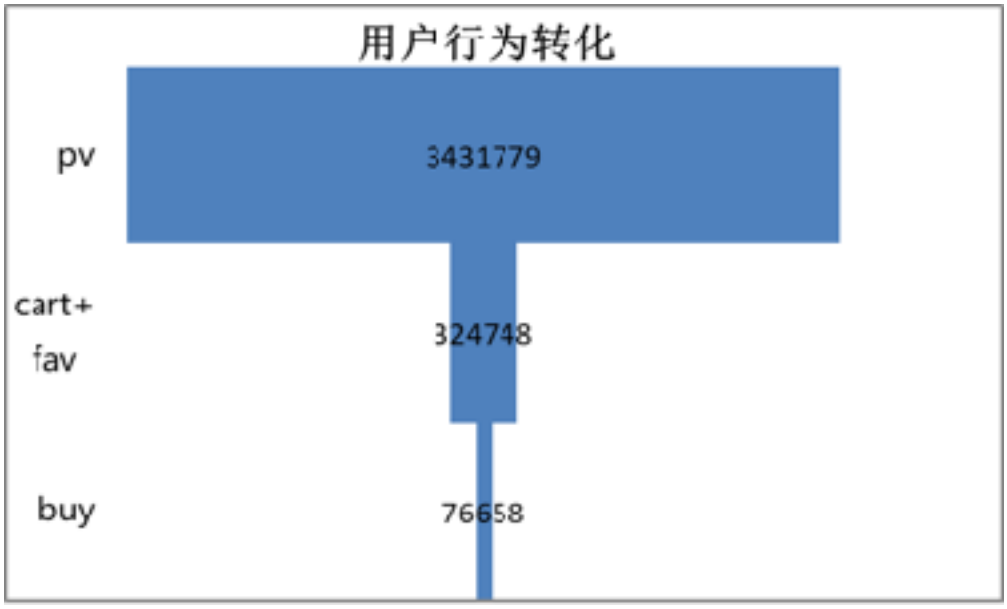
3.转化漏斗

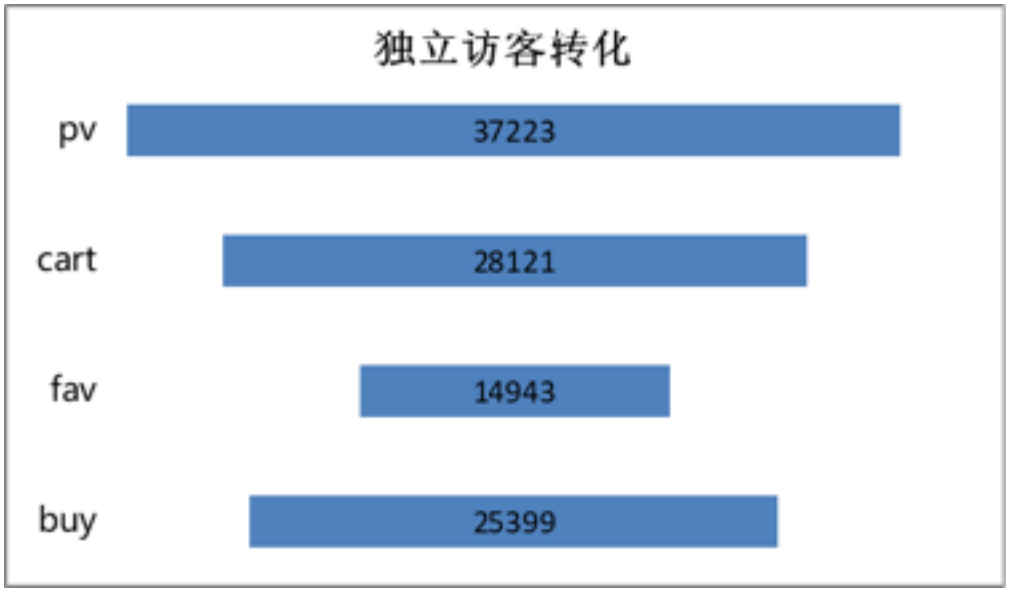
```
42  SELECT behavetype as 阶段,count(1) as 人数,count(distinct user_id)
    as 用户ID FROM user
43  group by behavetype
44  order by count(1) desc;
```

信息 | 结果 1 | 剖析 | 状态

阶段	人数	用户ID
pv	3431779	37223
cart	213608	28721
fav	111140	14943
buy	76658	25349

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%，若要了解具体环节的转化率，需要对整个流程接触到的版面进行细化的数据分析。

2、用户分析

（1）留存

次日留存回购人数：

```
SELECT t1.dates,COUNT(t1.dates),COUNT(t2.dates),concat(round(COUNT(
t2.dates)/count(t1.dates)*100,2),'%') as 留存率 from
(SELECT user_id,dates from user where behavetype='buy' group by
user_id,dates)t1 LEFT JOIN
(SELECT user_id,dates from user where behavetype='buy' group by
user_id,dates)t2
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	5378	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	5376	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：

```
SELECT t1.dates,COUNT(t1.dates),COUNT(t2.dates),concat(round(COUNT(
t2.dates)/count(t1.dates)*100,2),'%') as 留存率 from
(SELECT user_id,dates from user where behavetype='buy' group by
user_id,dates)t1 LEFT JOIN
(SELECT user_id,dates from user where behavetype='buy' group by
user_id,dates)t2
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	6661	0	0.00%
2017-12-01	5316	1355	25.49%
2017-11-30	5624	1352	24.04%
2017-12-03	6665	0	0.00%

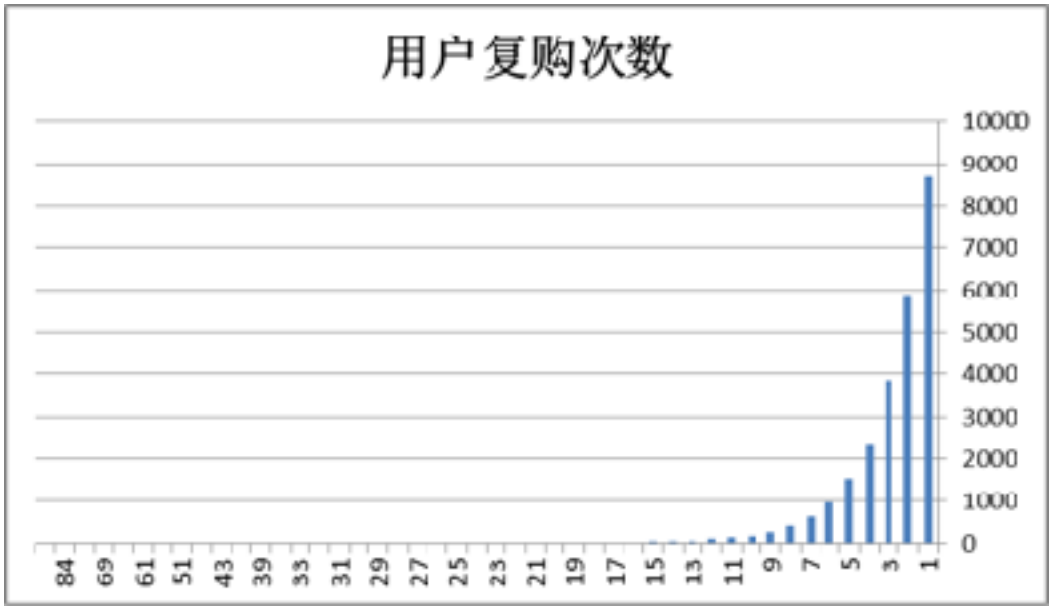
整体上看，次日留存率在22%-26%之间，3日留存率在22%-25%之间，说明在数据集时间范围内回购率还是很高的。

（2）复购：用户复购次数

```
SELECT user_id,count(*) from user
where behavetype='buy' GROUP BY user_id;
```

信息 | 结果 1 | 剖析 | 状态

user_id	count(*)
100	8
117	10



用户复购次数集中在1-5次，呈长尾分布，该阶段用户消费欲望不大。

（3）跳失率：只有浏览行为的用户数/总用户数

```
64  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
65  (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。

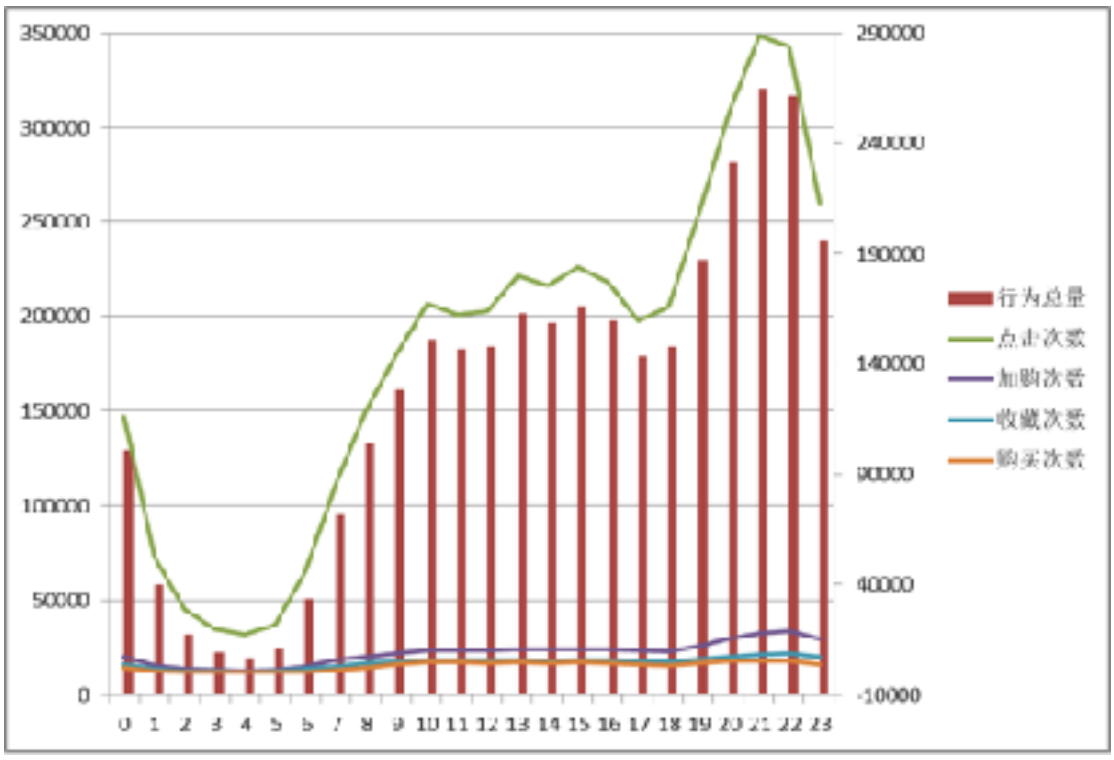
（4）用户消费习惯分析

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	20045	1310	706	239



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，该时间段处在下班、下课、吃晚饭的状态，符合用户生活习惯；

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。

（5）用户价值分析

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

```
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'
from user
where behavetype='buy'
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：

```
SELECT max(t.最后一次交易距今时间),max(t.交易次数) from
(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'
from user
where behavetype='buy'
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间)	max(t.交易次数)
10	84

2.对用户进行评分

```
SELECT user_id,(case WHEN t.最后一次交易距今时间 in (0,1) THEN 4
                    when t.最后一次交易距今时间 in (2,3) THEN 3
                    WHEN t.最后一次交易距今时间 in (4,5) THEN 2
                    WHEN t.最后一次交易距今时间 in (6,7) THEN 1
                    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
                      when t.交易次数 between 11 and 15 then 3
                      when t.交易次数 BETWEEN 6 and 10 then 2
                      when t.交易次数 BETWEEN 1 and 5 then 1
```

信息 | 结果 1 | 剖析 | 状态

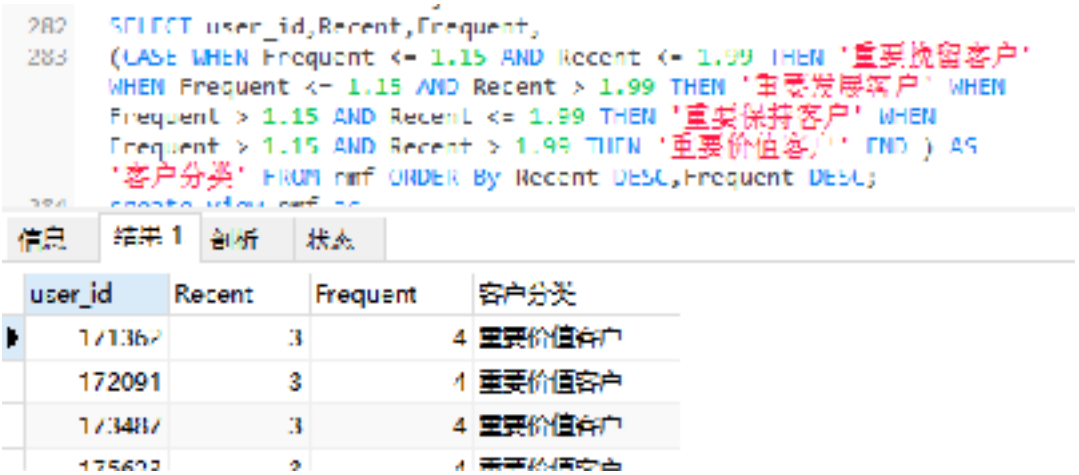
user_id	Recent	Frequent
220615	3	1
220625	3	1
220628	3	1

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：

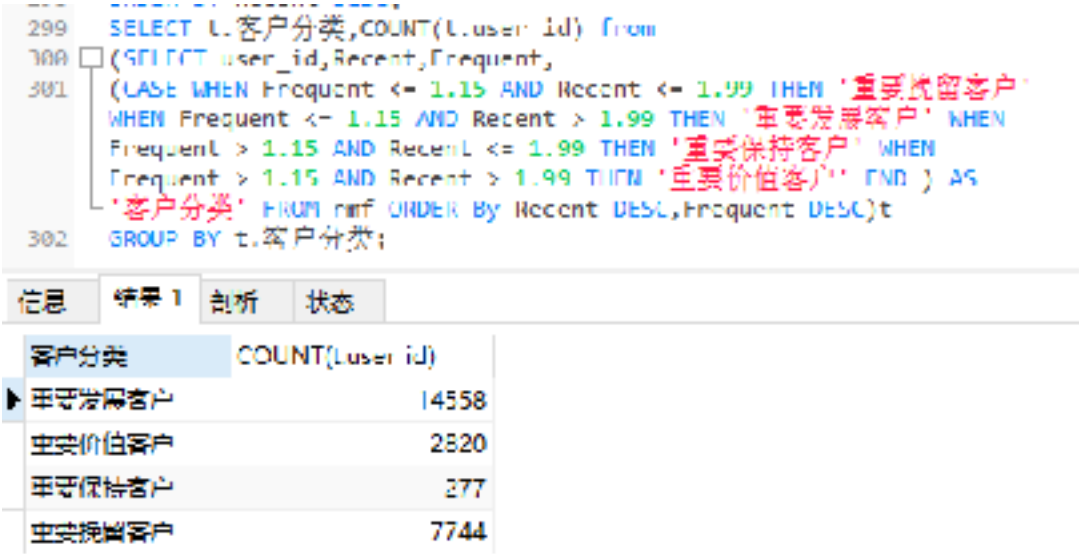
F（3-4）	重要保持客户	重要价值客户
F（1-2）	重要挽留客户	重要发展客户
	R（1-2）	R（3-4）

3.客户分层绘图

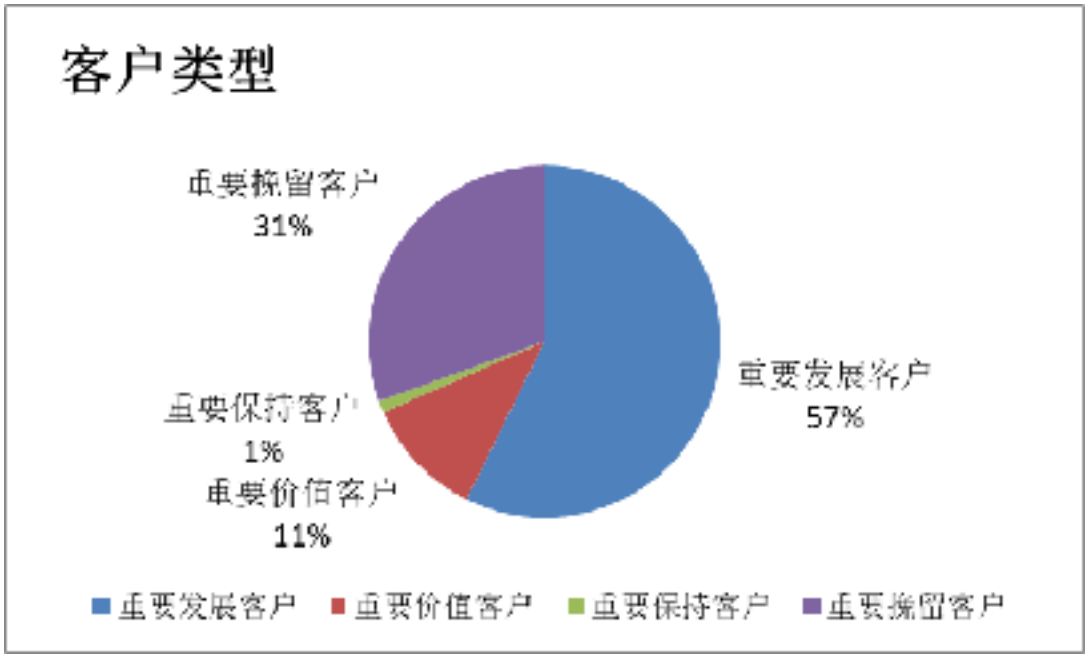


user_id	Recent	Frequent	客户分类
171362	3	4	重要价值客户
172091	3	4	重要价值客户
173487	3	4	重要价值客户

各类客户数量：



客户分类	COUNT(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



3、商品销售分析

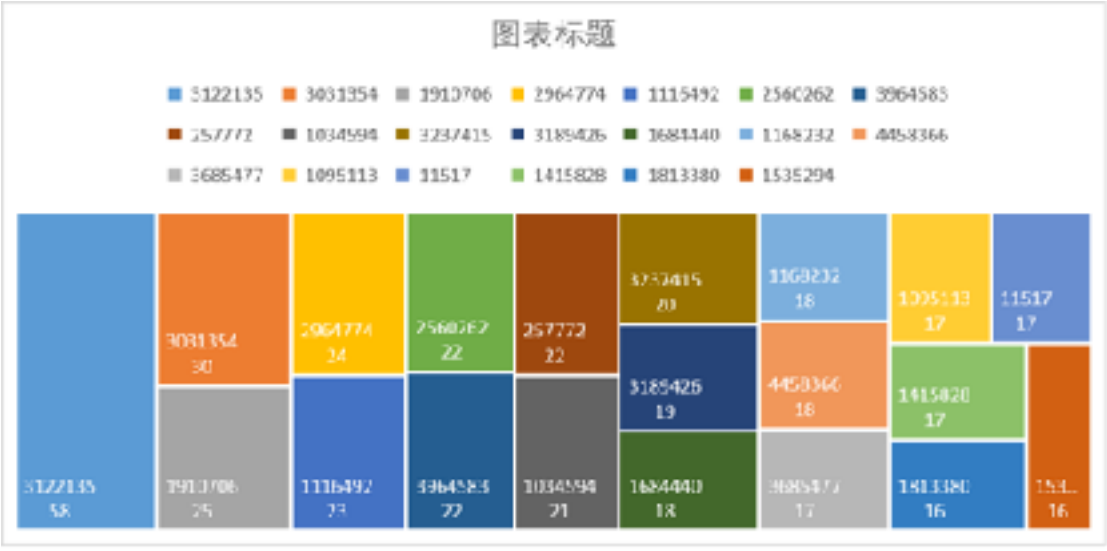
1）热销商品、热销类别

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

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

信息 | 结果 1 | 剖析 | 状态

item_id	ccount(DISTINCT user_id)
3122135	58
3031354	30

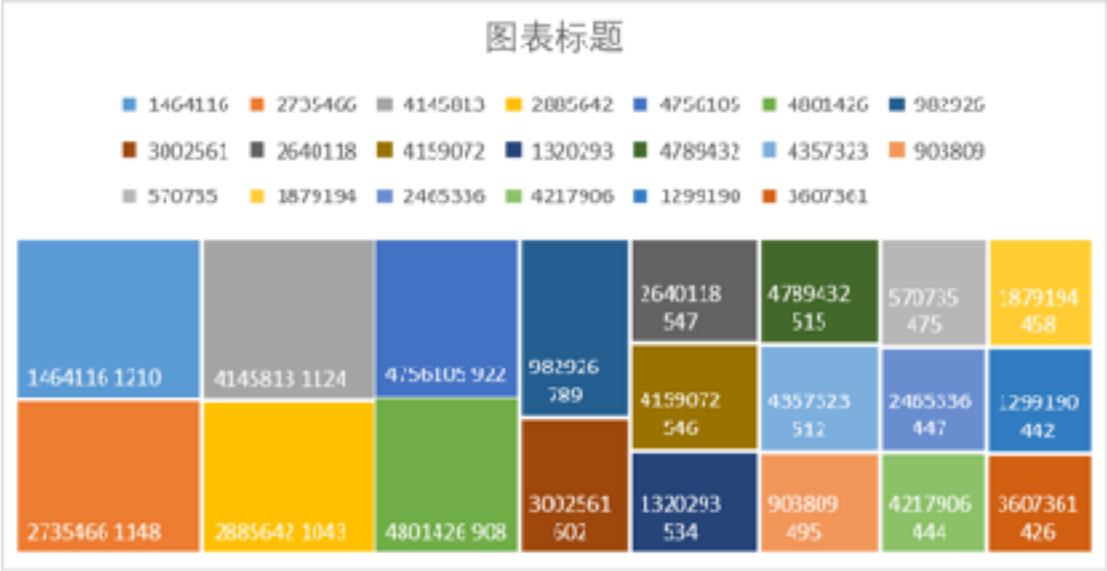


从类别上看，1464116标签号类别销量最好，共有1210个用户购买该类别商品：

```
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

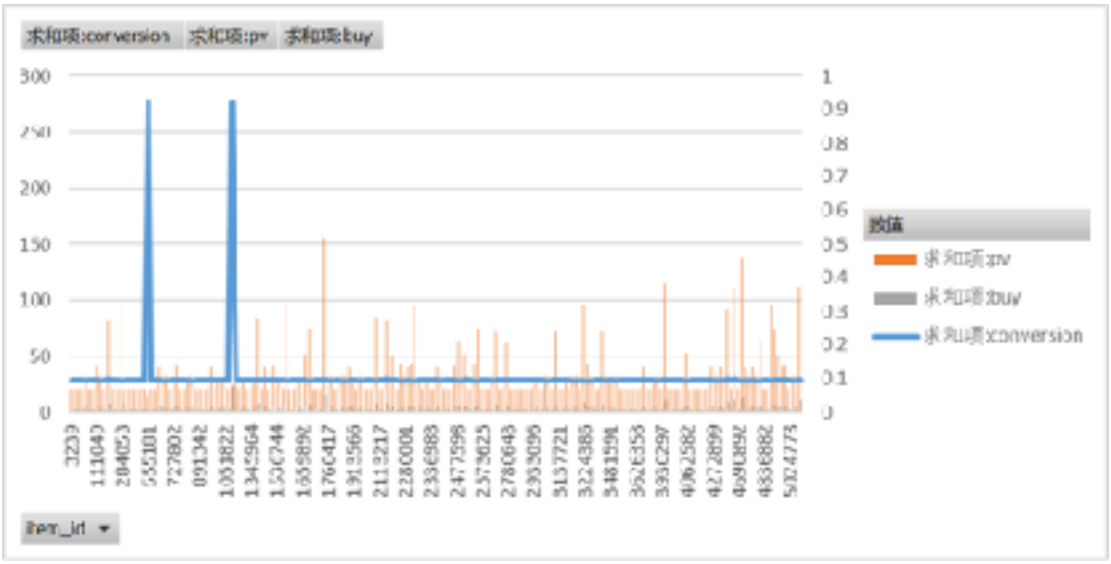


2）商品购买转化率

```
CREATE VIEW uniitem 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 uniitem
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%
4681900	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%，猜测可能是由于这两种商品在此期间进行了一系列促销活动导致的。对于其他商品浏览量高，而转化率非常低的，建议商家对商品购买流程使用转化漏斗进行分析并进行改善。