淘宝用户行为分析(SQL)

**#1.建库建表**

create database taobao;

use taobao;

drop table if exists tb;

数据来源于阿里天池，数据集包含了淘宝app2017年11月25-12月3日的用户行为数据，包括浏览、收藏、加购和购买。每一条数据代表用户的行为记录，包含用户编号user\_id,商品编号item\_id，商品类型category\_id，行为类型behavior\_type以及时间戳Time\_stamp。由于数据量上亿，文件太大，因此从中仅选用了100万样本。

行为类型包括四种：pv（浏览）、cart（加购）、fav（收藏）、buy（购买）

create table tb(

user\_id varchar(255) not null,

item\_id varchar(255),

category\_id varchar(255),

behavior\_type varchar(10),

Time\_stamp int(11)

);

#将csv文件导入数据库

load data infile 'F:/TianChi/OutPut/UserBehavior-009.csv'

into table tb

fields terminated by ','

optionally enclosed by '"'

escaped by '"'

lines terminated by '\r\n';

#复制一张表,以免丢失

create table ub select \* from tb;

select \* from ub;

SHOW DATABASES;

show tables;

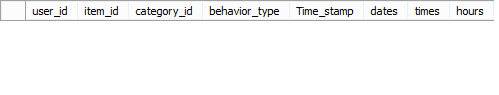
select count(\*) from tb;

**#2.查看是否有缺失值**

select \* from tb

where user\_id is null or item\_id is null or category\_id is null

or behavior\_type is null or Time\_stamp is null;#结果无缺失值



#另一种方法

select count(1) from tb where user\_id is null;

#查看是否有重复值

select count(\*) from tb

group by user\_id,item\_id,category\_id,behavior\_type,time\_stamp

having count(\*) >1;

Empty set (7.05 sec)

**#3.时间戳的处理**

SET SQL\_SAFE\_UPDATES = 0;

#(1)添加dates列，记录行为发生日期

alter table tb add dates varchar(255);

update tb set dates=from\_unixtime(Time\_stamp,'%Y-%m-%d');

#(2)添加时间列，记录发生时间

alter table tb add times varchar(255);

update tb set times =from\_unixtime(Time\_stamp,'%H:%i:%s');

**#4.异常值处理**

#(1)查看行为类型是否存在异常

select behavior\_type from tb

where behavior\_type not in ('pv','cart','buy','fav');

#(2)查看Timestamp是否存在异常

select dates from tb where dates not between '2017-11-25' and '2017-12-03';

#删除错误数据

delete from tb where dates not between '2017-11-25' and '2017-12-03';

#检查数据概况

select count(distinct user\_id) as '用户总数',

count(distinct item\_id) '商品总数',

count(distinct category\_id) '商品类型总数',

count(distinct behavior\_type) '行为类型总数',

count(distinct dates) as '天数'

from tb;

+----------+---------+-------------+-------------+------+

| 用户总数 | 商品总数 | 商品类型总数 | 行为类型总数 | 天数 |

+----------+---------+------------+-------------+------+

| 9739 | 398972 | 5793 | 4 | 9 |

+----------+----------+--------------+--------------+------+

#数据清洗完毕以后，进行分析和可视化

**#5.分析**

**#5.1分析用户每日行为规律**

#每日浏览量,用户数

select dates,count(behavior\_type) as '每日浏览量',count(distinct user\_id) as ‘每日浏览人数’

from tb

where behaviortype='pv'

group by dates

order by dates;

+------------+------------+--------------+

| dates | 每日浏览量 | 每日浏览人数 |

+------------+------------+--------------+

| 2017-11-25 | 93931 | 6782 |

| 2017-11-26 | 95657 | 6928 |

| 2017-11-27 | 87243 | 6828 |

| 2017-11-28 | 88638 | 6810 |

| 2017-11-29 | 91334 | 6930 |

| 2017-11-30 | 94735 | 7010 |

| 2017-12-01 | 98138 | 7054 |

| 2017-12-02 | 123514 | 9271 |

| 2017-12-03 | 122446 | 9300 |

+------------+------------+--------------+

9 rows in set (4 min 39.32 sec)

#每日加入购物车数量

+------------+------------+

| dates | 每日加购量 |

+------------+------------+

| 2017-11-25 | 5786 |

| 2017-11-26 | 5784 |

| 2017-11-27 | 5463 |

| 2017-11-28 | 5516 |

| 2017-11-29 | 5550 |

| 2017-11-30 | 5638 |

| 2017-12-01 | 6102 |

| 2017-12-02 | 7829 |

| 2017-12-03 | 7779 |

+------------+------------+

+------------+------------+

| dates | 每日收藏量 |

+------------+------------+

| 2017-11-25 | 2716 |

| 2017-11-26 | 3113 |

| 2017-11-27 | 2873 |

| 2017-11-28 | 2726 |

| 2017-11-29 | 3057 |

| 2017-11-30 | 2891 |

| 2017-12-01 | 2999 |

| 2017-12-02 | 3895 |

| 2017-12-03 | 3818 |

+------------+------------+

+------------+------------+

| dates | 每日购买量 |

+------------+------------+

| 2017-11-25 | 1974 |

| 2017-11-26 | 2022 |

| 2017-11-27 | 2229 |

| 2017-11-28 | 2220 |

| 2017-11-29 | 2299 |

| 2017-11-30 | 2323 |

| 2017-12-01 | 2151 |

| 2017-12-02 | 2536 |

| 2017-12-03 | 2605 |

+------------+------------+

9 rows in set (3.01 sec)

#每日购物用户占比

select a.dates,count(distinct b.user\_id)/count(distinct a.user\_id) as '每日购物用户占比'

from tb a

left join (select \* from tb where behavior\_type='buy') b

on a.user\_id=b.user\_id and a.item\_id=b.item\_id and a.time\_stamp=b.time\_stamp

group by a.dates;

+----------+-----------------+

| dates | 每日购物用户占比 |

+----------+-----------------+

| 2017-11-25 | 0.1888 |

| 2017-11-26 | 0.1857 |

| 2017-11-27 | 0.2014 |

| 2017-11-28 | 0.2005 |

| 2017-11-29 | 0.2054 |

| 2017-11-30 | 0.2038 |

| 2017-12-01 | 0.1950 |

| 2017-12-02 | 0.1804 |

| 2017-12-03 | 0.1840 |

+----------+-----------------+

#每日人均交易数

select a.dates,count(b.user\_id)/count(distinct a.user\_id) as '每日人均交易数'

from tb a

left join (select \* from tb where behavior\_type='buy') b

on a.user\_id=b.user\_id and a.item\_id=b.item\_id and a.time\_stamp=b.time\_stamp

group by a.dates;

+------------+----------------+

| dates | 每日人均交易数 |

+----------+---------------+

| 2017-11-25 | 0.2830 |

| 2017-11-26 | 0.2837 |

| 2017-11-27 | 0.3173 |

| 2017-11-28 | 0.3157 |

| 2017-11-29 | 0.3219 |

| 2017-11-30 | 0.3207 |

| 2017-12-01 | 0.2959 |

| 2017-12-02 | 0.2651 |

| 2017-12-03 | 0.2725 |

+----------+---------------+

11月25-11月26与12月2日-12月3日均为周末，相比前一个周末，12月的周末浏览量和加购量明显更多，但最终的购买量没有明显增加，推测有可能是双十二活动预热和服饰焕新所致。

注：

* 母婴冬季保暖节（11/20-11/30）
* 黑色星期五（11/22-11/26）
* 咖啡节（11/23-11/27）
* 火拼节（11/27-11/29）
* “双十二”预热阶段（12/1-12/8）
* 服饰焕新（12/1-12/6）

#从每日不同时段分析用户行为规律：增加一个字段’hours’

select hours,count(\*) as '总浏览量' from tb

where behavior\_type='pv'

group by hours order by hours;

+-------+----------+

| hours | 总浏览量 |

+-------+----------+

| 00 | 30396 |

| 01 | 13852 |

| 02 | 7984 |

| 03 | 5636 |

| 04 | 4969 |

| 05 | 6060 |

| 06 | 12071 |

| 07 | 22225 |

| 08 | 30016 |

| 09 | 37185 |

| 10 | 43649 |

| 11 | 42704 |

| 12 | 42168 |

| 13 | 46453 |

| 14 | 45928 |

| 15 | 47250 |

| 16 | 46880 |

| 17 | 41418 |

| 18 | 42911 |

| 19 | 55026 |

| 20 | 65406 |

| 21 | 75030 |

| 22 | 74687 |

| 23 | 55732 |

+------+---------+

#总加购量

select hours,count(\*) as '总加购量' from tb

where behavior\_type='cart'

group by hours order by hours;

+-------+----------+

| hours | 总加购量 |

+-------+----------+

| 00 | 1824 |

| 01 | 892 |

| 02 | 522 |

| 03 | 372 |

| 04 | 267 |

| 05 | 374 |

| 06 | 855 |

| 07 | 1499 |

| 08 | 1869 |

| 09 | 2323 |

| 10 | 2666 |

| 11 | 2663 |

| 12 | 2538 |

| 13 | 2634 |

| 14 | 2730 |

| 15 | 2754 |

| 16 | 2867 |

| 17 | 2574 |

| 18 | 2459 |

| 19 | 3241 |

| 20 | 3922 |

| 21 | 4649 |

| 22 | 4899 |

| 23 | 4054 |

+-------+----------+

24 rows in set (3.08 sec)

#总收藏量

select hours,count(\*) as '总收藏量' from tb

where behavior\_type='fav'

group by hours order by hours;

+-------+------------+

| hours | 总收藏量 |

+-------+------------+

| 00 | 956 |

| 01 | 438 |

| 02 | 266 |

| 03 | 180 |

| 04 | 160 |

| 05 | 217 |

| 06 | 490 |

| 07 | 689 |

| 08 | 988 |

| 09 | 1259 |

| 10 | 1418 |

| 11 | 1384 |

| 12 | 1333 |

| 13 | 1484 |

| 14 | 1342 |

| 15 | 1381 |

| 16 | 1479 |

| 17 | 1429 |

| 18 | 1294 |

| 19 | 1529 |

| 20 | 1879 |

| 21 | 2172 |

| 22 | 2374 |

| 23 | 1947 |

+-------+------------+

24 rows in set (3.04 sec)

#总购买量

select hours,count(\*) as '总购买量' from tb

where behavior\_type='buy'

group by hours order by hours;

+-------+----------+

| hours | 总购买量 |

+-------+---------+

| 00 | 576 |

| 01 | 236 |

| 02 | 148 |

| 03 | 65 |

| 04 | 88 |

| 05 | 73 |

| 06 | 155 |

| 07 | 388 |

| 08 | 643 |

| 09 | 970 |

| 10 | 1279 |

| 11 | 1272 |

| 12 | 1264 |

| 13 | 1334 |

| 14 | 1178 |

| 15 | 1214 |

| 16 | 1252 |

| 17 | 1081 |

| 18 | 898 |

| 19 | 1181 |

| 20 | 1355 |

| 21 | 1380 |

| 22 | 1367 |

| 23 | 962 |

+-------+------------+

24 rows in set (2.66 sec)

晚上19：00-22：00各项行为总量呈现明显上升趋势，21：00-22：00用户行为总量达到顶峰状态，22：00至次日凌晨4点，各项用户行为数量均呈现明显下降趋势，多项用户行为于次日凌晨4时达到最低。晚上19：00过后为多数用户休息时间，从一天工作中解放出来，有时间浏览购物，凌晨为多数用户睡眠时间，因此用户行为总量降低。

**#5.2用户留存分析**

对从11月25日开始到12月3期间用户的新增与留存分析。首日（11月25日）的新用户为当天所有存在任意行为的用户，其余日期的新用户定义为此前未出现任何行为的用户，而当天存在任何行为的用户。

#找出第一次使用的用户与日期

select user\_id,min(dates) as firstday from tb

group by user\_id;

#用户所有使用时间

select user\_id,dates from tb

group by user\_id,dates;

#将用户id、使用时间和首次使用时间放在一张虚拟表

create view v\_time as

select a.user\_id,a.dates,b.firstday

from (select user\_id,dates from tb group by user\_id,dates) a

join

(select user\_id,min(dates) as firstday from tb group by user\_id) b

on a.user\_id=b.user\_id

order by a.user\_id,a.dates;

#获取第一次使用时间和后续使用时间间隔

create view v\_diff as

select user\_id,dates,firstday,datediff(dates,firstday) as diff\_day from v\_time;

#计算留存日的用户数量

create view cus\_qua as

select firstday,

sum(case when diff\_day=0 then 1 else 0 end) as '当日新增用户',

sum(case when diff\_day=1 then 1 else 0 end) as '1日后留存用户',

sum(case when diff\_day=2 then 1 else 0 end) as '2日后留存用户',

sum(case when diff\_day=3 then 1 else 0 end) as '3日后留存用户',

sum(case when diff\_day=4 then 1 else 0 end) as '4日后留存用户',

sum(case when diff\_day=5 then 1 else 0 end) as '5日后留存用户',

sum(case when diff\_day=6 then 1 else 0 end) as '6日后留存用户',

sum(case when diff\_day=7 then 1 else 0 end) as '7日后留存用户',

sum(case when diff\_day=8 then 1 else 0 end) as '8日后留存用户'

from v\_diff

group by firstday order by firstday;

#计算留存率(round(\*,3)四舍五入，保留3位小数)

select firstday,当日新增用户,

concat(round(1日后留存用户/当日新增用户,3)\*100,'%') as '1日留存率',

concat(round(2日后留存用户/当日新增用户,3)\*100,'%') as '2日留存率',

concat(round(3日后留存用户/当日新增用户,3)\*100,'%') as '3日留存率',

concat(round(4日后留存用户/当日新增用户,3)\*100,'%') as '4日留存率',

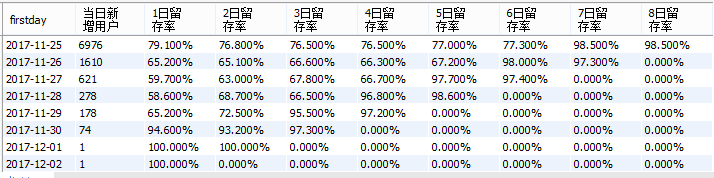
concat(round(5日后留存用户/当日新增用户,3)\*100,'%') as '5日留存率',

concat(round(6日后留存用户/当日新增用户,3)\*100,'%') as '6日留存率',

concat(round(7日后留存用户/当日新增用户,3)\*100,'%') as '7日留存率',

concat(round(8日后留存用户/当日新增用户,3)\*100,'%') as '8日留存率'

from cus\_qua order by firstday;



由于没有前置的用户数据，因此得到的新增用户与实际略有差异，但观察新增数据发现，每日新增用户呈现下降趋势，并且留存率在58%以上，12月2日和3日的留存率最高，符合双十二预热带来的用户关注度。

#用户购买情况

#这里主要从复购率角度，复购率指的是产生两次或者两次以上的购买的用户占购买用户的占比

#每位用户购买次数

create view rebuy as

select user\_id,count(behavior\_type) as buy\_times

from tb

where behavior\_type='buy'

group by user\_id

having count(behavior\_type)>=2

order by buy\_times desc;

#计算复购率

select count(distinct user\_id) from rebuy;#计算购买次数两次及以上的用户数量

select count(distinct user\_id) from tb

where behavior\_type='buy';#所有购买行为的用户数

select (select count(distinct user\_id) from rebuy)/(select count(distinct user\_id)from tb

where behavior\_type='buy') as '复购率';

C:\Users\lenovo\AppData\Local\Temp\1618300791(1).png

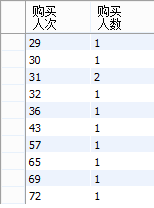
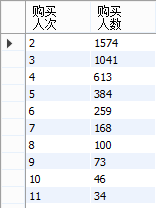
#复购用户中购买人次，人数

select buy\_times as '购买人次',count(distinct user\_id) as '购买人数'

from rebuy

group by 购买人次

order by 购买人次 desc;



从上面可以看到，11/25-12/3大部分复购用户的购买次数集中在2-6次，有9位用户在这9天的购买次数达到30次以上，最高达到了72次；说明淘宝用户忠诚度较高。

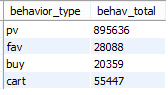
**#5.3用户行为路径分析**

#不同行为转化率

select behavior\_type,count(behavior\_type) as behav\_total

from tb

group by behavior\_type;



从转化漏斗可以看出，用户从点击到加购的转化率为6.19%，从点击到收藏的转化率为3.14%，

从点击到购买的转化率为2.27%。

**#转化率低的原因？**

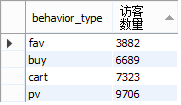
* 假设1：用户只是浏览商品而没有购买商品？
* 假设2：用户加购/收藏到购买的转化率更高？

#假设1：用户只是浏览商品而没有购买商品

#访客行为转化率

select behavior\_type,count(distinct user\_id) '访客数量'

from tb group by behavior\_type;



从pv到buy，转化率为68.92%,远远高于上一层行为漏斗的2.27%，因此假设1不成立。

#

#假设2：用户加购/收藏到购买的转化率高

#创建视图，统计不同用户不同商品行为:比较用户从浏览到购买与从收藏/加购到购买的转化率

create view buy\_way as

select user\_id,item\_id,

sum(if(behavior\_type='pv',1,0)) as pview,

sum(if(behavior\_type='cart',1,0)) as cart,

sum(if(behavior\_type='fav',1,0)) as favor,

sum(if(behavior\_type='buy',1,0)) as buy

from tb group by user\_id,item\_id;

#a.浏览总行为数

select count(\*) as '总浏览量' from buy\_way where pview>0;

#b.浏览到流失

select count(\*) as '浏览-流失'

from buy\_way

where pview>0 and cart=0 and favor=0 and buy=0;

#浏览-收藏行为计数

select count(\*) as '浏览-收藏' from buy\_way

where pview>0 and favor>0 and cart=0;

#浏览-加购行为

select count(\*) as '浏览-加购' from buy\_way

where pview>0 and cart>0 and favor=0;

#浏览-购买行为

select count(\*) as '浏览-购买' from buy\_way

where pview>0 and buy>0 and favor=0 and cart=0;

#流失

#浏览-收藏-流失

select count(\*) as '浏览-收藏-流失'

from buy\_way

where pview>0 and favor>0 and cart=0 and buy=0;

#浏览-加购-流失

select count(\*) as '浏览-加购-流失'

from buy\_way

where pview>0 and favor=0 and cart>0 and buy=0;

#浏览-收藏-购买

select count(\*) as '浏览-收藏-购买'

from buy\_way

where pview>0 and fav>0 and buy>0 and cart=0;

#浏览-加购-购买

select count(\*) as '浏览-加购-购买'

from buy\_way

where pview>0 and fav=0 and buy>0 and cart>0;

#浏览-收藏&加购

select count(\*) as '浏览-收藏&加购'

from buy\_way

where pview>0 and fav>0 and cart>0;

#浏览-收藏&加购-购买

select count(\*) as '浏览-收藏&加购-购买'

from buy\_way

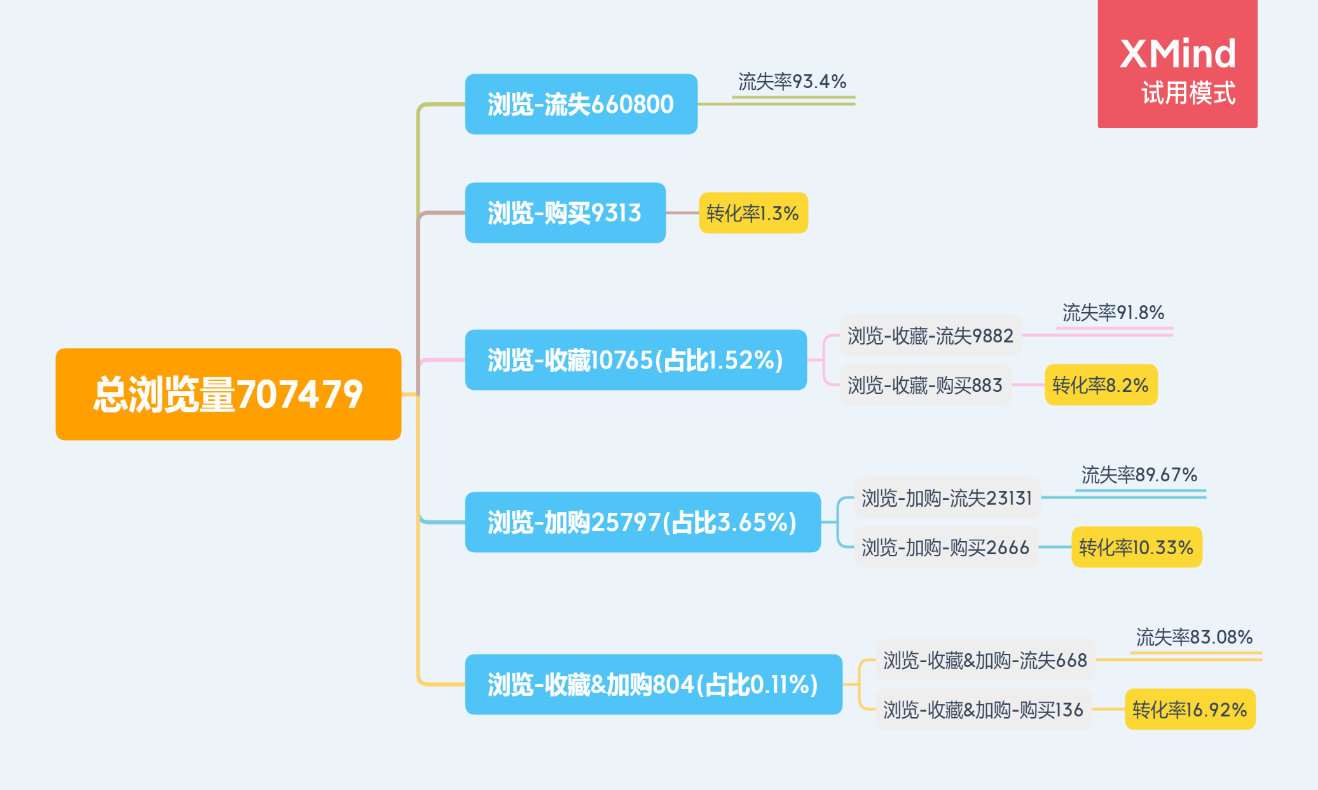
where pview>0 and fav>0 and buy>0 and cart>0;

#浏览-收藏&加购-流失

select count(\*) as '浏览-收藏&加购-流失'

from buy\_way

where pview>0 and fav>0 and buy=0 and cart>0;



从上图可知，93.4%的用户点击后便流失了，浏览后直接购买的占比仅为1.3%，浏览后加购的比例则有3.65%，高于浏览后收藏的比例（1.52%）。从购买转化率来看，收藏的购买转化率为8.2%，加购的购买转化率更高，为10.33%，说明用户收藏的商品可能只是喜欢，但还没有很大的购买意向。从这里，可以针对性做一些措施增加用户的购买意向，如发放优惠券或者给予商品一定折扣优惠。收藏并加购的购买转化率比较高，达到16.92%，这里可以做一些营销手段引导用户收藏并加购。

**#5.4消费者偏好分析**

#按商品分组计算其成交次数

select item\_id,count(\*) as tra\_amou

from tb

where behavior\_type='buy'

group by item\_id order by tra\_amou desc;

#按成交次数分组计算商品数量(如成交x次的商品有多少个)

select t.tra\_amou,count(\*)

from (select item\_id,count(\*) as tra\_amou

from tb

where behavior\_type='buy'

group by item\_id) t

group by t.tra\_amou order by t.tra\_amou;

从图可知，15536个商品的成交次数仅为1次，占比为88.45%的，说明了大部分的成交商品为长尾商品，并没有看出特别带动销量的爆款。为了提高销量，商品页可以将畅销品和非畅销品展示在一起，或者捆绑销售，或者将爆款产品集中推送，提高整个平台的销量。

##筛选销量前10的商品

select item\_id,

sum(if(behavior\_type='buy',1,0)) as '商品成交次数',

sum(if(behavior\_type='pv',1,0)) as '商品点击量',

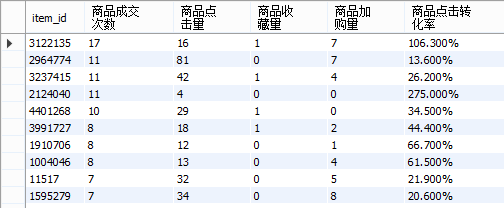
sum(if(behavior\_type='fav',1,0)) as '商品收藏量',

sum(if(behavior\_type='cart',1,0)) as '商品加购量',

concat(round(sum(if(behavior\_type='buy',1,0))/sum(if(behavior\_type='pv',1,0)),3)\*100,'%')

as '商品点击转化率' from tb

group by item\_id order by 商品成交次数 desc;



由于数据样本不够大，商品点击率略有不合理，但是不影响后续分析。

**#卖的好的商品的原因？**

* 假设1：流量多
* 假设2：转化率高？收藏量高？加购量高？
* 假设3：复购率高？

#（1）流量高？

按商品浏览量排序

select item\_id,

sum(if(behavior\_type='buy',1,0)) as '商品成交次数',

sum(if(behavior\_type='pv',1,0)) as '商品点击量',

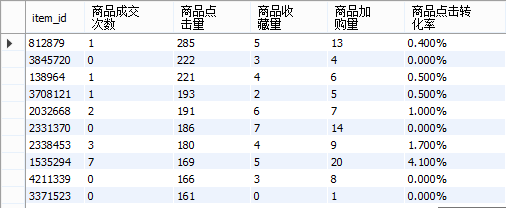
sum(if(behavior\_type='fav',1,0)) as '商品收藏量',

sum(if(behavior\_type='cart',1,0)) as '商品加购量',

concat(round(sum(if(behavior\_type='buy',1,0))/sum(if(behavior\_type='pv',1,0)),3)\*100,'%') as '商品点击转化率'

from tb

group by item\_id order by 商品点击量 desc;



从商品流量角度来看，浏览量高的商品并未出现在销量最高的商品之中，因此可见商品流量并非商品购买的主要因素，因此假设1不成立。

#（2）转化率高？即收藏量or加购量高？

#收藏量高？

select item\_id,

sum(if(behavior\_type='buy',1,0)) as '商品成交次数',

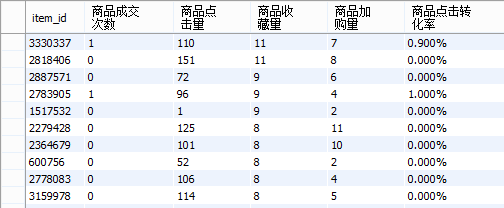
sum(if(behavior\_type='pv',1,0)) as '商品点击量',

sum(if(behavior\_type='fav',1,0)) as '商品收藏量',

sum(if(behavior\_type='cart',1,0)) as '商品加购量',

concat(round(sum(if(behavior\_type='buy',1,0))/sum(if(behavior\_type='pv',1,0)),3)\*100,'%') as '商品点击转化率' from tb

group by item\_id order by 商品收藏量 desc;



显然，收藏量高的商品并未出现在销量最高商品当中，侧面印证了商品收藏后的购买转化率较低；

#加购量高？

select item\_id,

sum(if(behavior\_type='buy',1,0)) as '商品成交次数',

sum(if(behavior\_type='pv',1,0)) as '商品点击量',

sum(if(behavior\_type='fav',1,0)) as '商品收藏量',

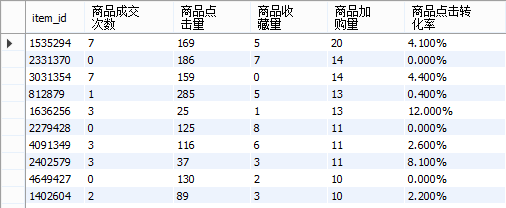
sum(if(behavior\_type='cart',1,0)) as '商品加购量',

concat(round(sum(if(behavior\_type='buy',1,0))/sum(if(behavior\_type='pv',1,0)),3)\*100,'%')

as '商品点击转化率'

from tb

group by item\_id order by 商品加购量 desc;



商品加购量高的有两个商品成交次数为7，属于并排销量前10的商品，说明用户加购后的购买转化率高于收藏后的购买转化率。

#（3）复购率高？

#筛选高平均成交量的top10商品

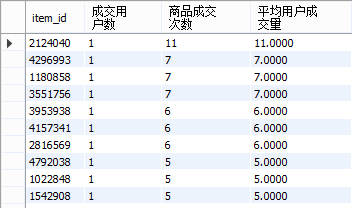
select item\_id,count(distinct user\_id) as '成交用户数',

sum(if(behavior\_type='buy',1,0)) as '商品成交次数',

sum(if(behavior\_type='buy',1,0))/count(distinct user\_id) as '平均用户成交量'

from tb where behavior\_type='buy'

group by item\_id order by 平均用户成交量 desc;



显然，有几个商品的与销量高商品重合，所以私以为假设3成立。由图可知，个别商品成交用户数为1，但成交次数达到11次，可见老客户复购率的提升能有效提高销量和营收。

**附：**

由于100万样本仍有些不足，因此在商品偏好分析这块增加到200万样本来做一下对比，得出的结论和100万样本相似。

