淘宝用户点击行为数据分析报告

对淘宝 2022 年 11 月 24 日-2022 年 12 月 3 日用户行为数据进行分析

李梓昊 2023 年 9 月 12 日

一．数据介绍

本数据来源于阿里巴巴提供的淘宝用户行为数据集，主要包含了 2022 年 11 月 5 日到 2022 年 12 月 3 日之间，共计 1993975 条随机用户的数据。

列名	#	数据类型	非空	自增
123 user	1	int	[ ]	[ ]
123 item	2	int	[ ]	[ ]
123 category	3	int	[ ]	[ ]
ABC behavior	4	varchar(50)	[ ]	[ ]
123 time	5	int	[ ]	[ ]

代码 1 数据类型展示

本数据共有 5 个字段，分别如下：

字段名	解释
user	用户编号 int
item	商品类目编号 int
category	商品类目编号 int
behavior	用户行为 varchar(50)
time	用户操作时间 int

二．数据清洗与处理

1. 通过查询语句，确认数据中无缺失值。
2. 为方便接下来的指标计算与分析，将时间数据转化为标准的“年-月-日”格式和对应的当天小时数。

```
create table tb as
select
    user, item, category,
    behavior, FROM_UNIXTIME(time,'%Y-%m-%d') date,
    from_unixtime(time,'%H') time
from UserBehavior;
```

代码 2 时间类型数据处理

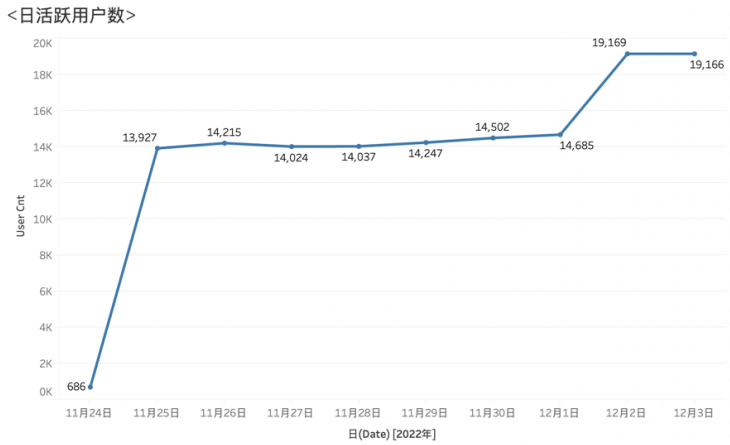
三．数据分析部分

1. 用户活跃度分析

```
#日活量
select date, COUNT(DISTINCT user) user_cnt
from taobao
group by date
order by date asc;
```

代码 3 用户日活量计算

通过源数据按日期分组操作，并按日期升序排序，计算了各个用户出现的次数，由此得到每日用户活跃人数。在此去除了数据异常点 2022-11-24 日之前的数据。



图一 日活跃用户数

数据描述：由折线图中可以看出从 2022 年 11 月 24 日到 2022 年 12 月 3 日期间，每日的用户活跃人数整体呈上升趋势；但在 2022 年 11 月 25 日到 2022 年 12 月 1 日期间用户活跃数基本持平，直到 2022 年 12 月 2 日才出现大幅增长。

推测原因是用户在经历了淘宝“双十一“活动后，整体消费热情有所冷却，但随着 2022 年 12 月 1 日进入淘宝”双十二“又一网购活动，用户带着消费热情，或本着”临近活动，进来转转“的想法又一次点进了淘宝，由此带来了大幅用户活跃度的提升。

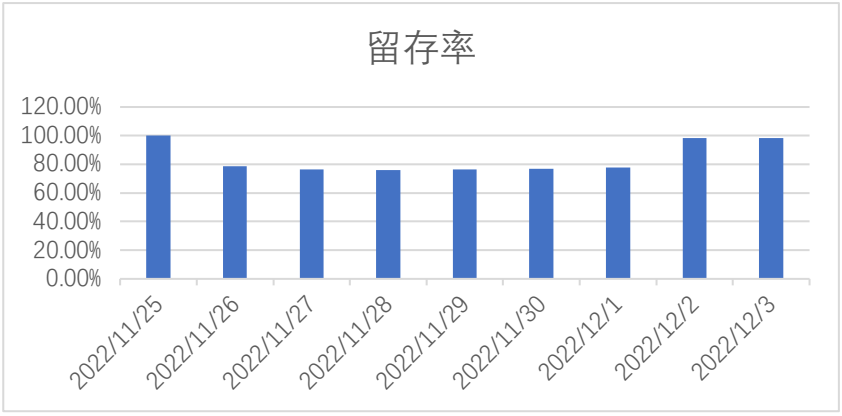
由此可见，淘宝每年的“双十一“与”双十二“活动吸引用户的能力很强。

2. 用户留存率分析

```
#留存计算
with tmp1 as (select DISTINCT user, date, min(date) over(partition by user) fd,
datediff(date, min(date) over(partition by user)) by_day
from tb)
select fd,
sum(case when by_day = 0 then 1 else 0 end) as day0,
sum(case when by_day = 1 then 1 else 0 end) as day1,
sum(case when by_day = 2 then 1 else 0 end) as day2,
sum(case when by_day = 3 then 1 else 0 end) as day3,
sum(case when by_day = 4 then 1 else 0 end) as day4,
sum(case when by_day = 5 then 1 else 0 end) as day5,
sum(case when by_day = 6 then 1 else 0 end) as day6,
sum(case when by_day = 7 then 1 else 0 end) as day7,
sum(case when by_day = 8 then 1 else 0 end) as day8
from tmp1
group by fd
order by fd
```

代码 4 用户留存数计算

按 user 分组，计算出各用户最早活跃日期和与该日期相差天数小于 7 天的所有日期，并按日期加和。



图二 用户留存率

数据描述：由图中可得从 2022 年 11 月 24 日到 2022 年 12 月 3 日期间用户留存率基本与用户活跃数呈相同态势，11 月 25 日登录的用户在接下来七天内的留存率较稳定，接着在 12 月 2 日和 12 月 3 日留存率大幅上升，该变化可能是因为“双十二”活动所致。

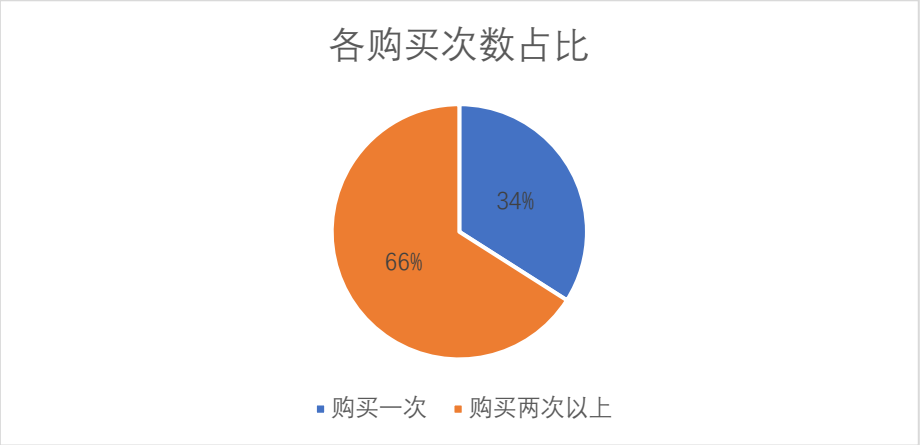
3. 转化及购买次数

```sql
#转化（购买）
#购买人数
select count(DISTINCT user)
from taobao
where behavior = 'buy' #输出结果是13330
```

代码 5 计算转化（购买）人数

选择 behavior=”buy”的不同用户，进行 count 计数。

图三 购买次数占比



数据描述：由图可知，购买一次的用户人数占比仅为 34%，购买两次以上的占比近 2/3，这说明复购率较高。同时，通过计算总购买次数与总购买用户数可以得到平均的购买次数为：

$$\frac{总购买次数}{总购买用户数}=\frac{40243}{13330}\approx 3.02$$

相对于 66%的复购率来讲，这样 3 左右的平均购买次数并不算高。

因此综合以上分析和平均购买次数来讲，说明淘宝对用户的吸引力很不错，“双十一”等活动都可以为淘宝的流量带来显著的提；同时，在日常中无活动期间，淘宝也能够保持较高的用户留存率，说明用户黏性较强；近 2/3 的复购率也说明用户对淘宝的认可度较高，但在 2022 年 11 月 24 日到 2022 年 12 月 3 日期间的平均购买次数并不算高，这也许是由于“双十一“刚过不久，用户的消费热情有所退却，或是临近月底用户等待发工资，这需要进一步与往年数据进行纵向对比寻求原因。

4. 用户漏斗

```sql
# 收藏数
select count(user)
from taobao
where behavior = 'cart' #111015

# 加购数
select count(user)
from taobao
where behavior = 'fav' #57526

# 总数
select count(user)
from taobao
where behavior = 'pv' #1791216

# 购买数
select count(user)
from taobao
where behavior = 'buy' #40243
```

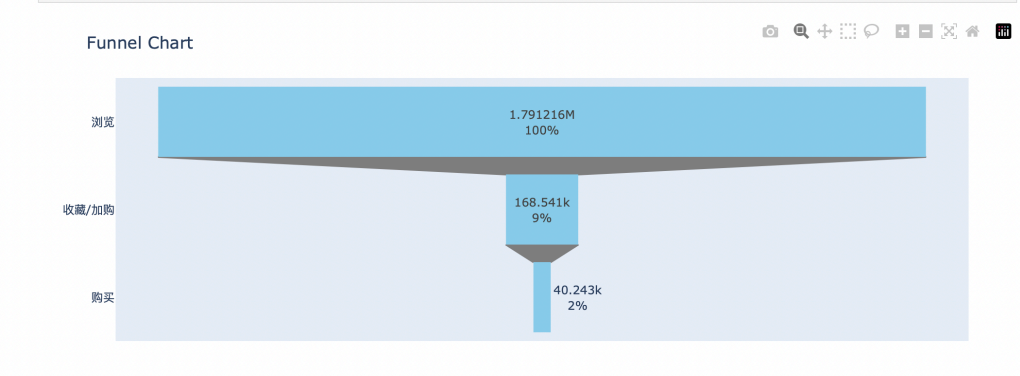
代码 6 计算用户漏斗中各指标

分别选取 behavior = cart、fav、pv、buy 这四类行为的数据，并进行 count 计数。

标签名	总数
访客数	19544
pv1	1
pv	1791216
Cart/fav	111015/57526
buy	40203

表二 跳失率 各环节转化率

图四 用户漏斗



数据描述：由以上数据可知，首先，淘宝的跳失率是很低的，这说明淘宝的首页很有吸引力，同时产品也是比较能够满足客户需求。

通过用户漏斗可以看到，各环节，无论是浏览——加购/收藏，还是加购/收藏——购买的转化率都不算低，这说明，首先，用户们都比较了解界面操作，其次，对于加购和收藏这两个功能相似的操作来讲，用户更愿意使用加购，这可能与界面中加购的图标更明显有关；另外，这样的高转化率和高复购率结合较低的平均购买次数来看，说明淘宝能够吸引大批用户浏览和加购，但产品力需要加强。

因此，我们可以增加浏览量或曝光量，由此扩大用户的点击量，进而增加转化量；同时也应注重产品力的提升，增强用户购买的意愿。

5.时序分析

（1）一周内用户行为趋势变化

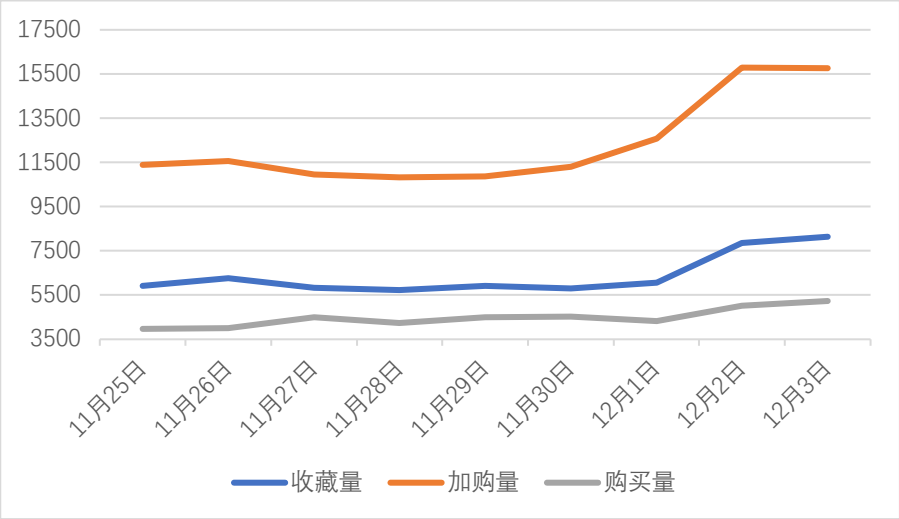
```sql
#用户行为与在不同时间段下的变化规律（时序分析）
#以2022.11.25-2022.12.03这九天为例
select
    date,
    sum(case when behavior='pv' then 1 else 0 end) 浏览,
    sum(case when behavior='fav' then 1 else 0 end) 收藏,
    sum(case when behavior='cart' then 1 else 0 end) 购物车,
    sum(case when behavior='buy' then 1 else 0 end) 购买
from taobao
where date BETWEEN '2022-11-25' AND '2022-12-03'
group by date
order by date;  #可以做个相关性分析
```

代码 7 2022-11-25 至 2022-12-03 中用户各行为数计算

选取 2022-11-25 至 2022-12-03 这九天的数据按日期划分，计算各用户各行为数。

图五 一周内浏览量变化

图六 一周内收藏/加购/购买量的变化



数据描述：根据上图，我们可以看到收藏量和加购量整体的走势基本同浏览量的走势一致。但购买量在某些地方却有些差别，比如 11 月 26 日至 11 月 27 日，购买量反而上升，这也许是因为周六日的影响；同时，从 12 月 1 日开始，加购和收藏量大幅上升，购买量也有所增加，这可能是因为开月发工资，加上清理一些之前加购或收藏的商品；但购买量无太多增加，这也许是因为用户在等待“双十二”活动，于是先将意向商品加购或收藏，造成加购和收藏的大幅上升。

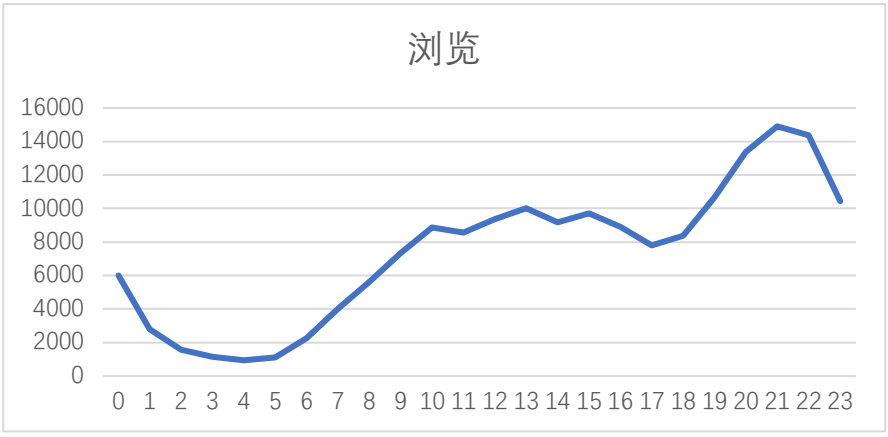
（2）一天内用户行为变化趋势

```sql
# 选取11.27周日这天做分析
select
    time,
    sum(case when behavior='pv' then 1 else 0 end) 浏览,
    sum(case when behavior='fav' then 1 else 0 end) 收藏,
    sum(case when behavior='cart' then 1 else 0 end) 购物车,
    sum(case when behavior='buy' then 1 else 0 end) 购买
from taobao
where date = '2022-11-27'
group by time
order by time;  #可以做个相关性分析
```

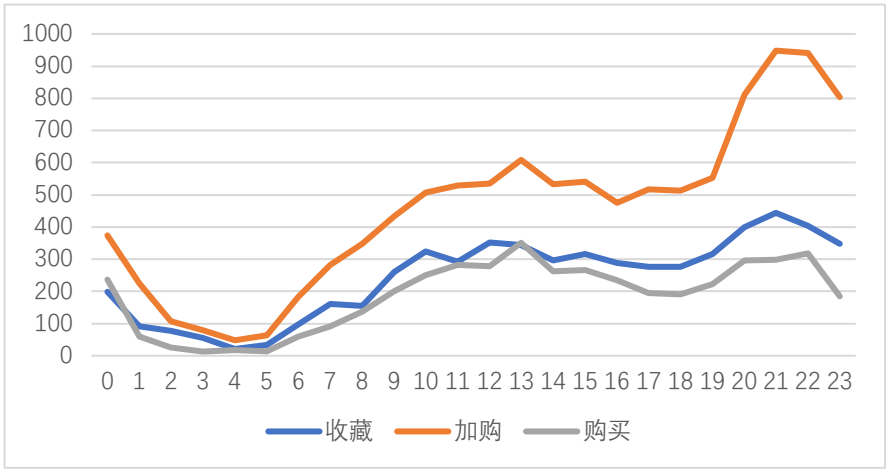
代码 8 选取 2022-11-27 日各时段用户行为数

选取 2022-11-27 日这天的数据，并按时段分组计算各行为的总数。

图七 一天内浏览量的变化



图八 一天内收藏/加购/购买量的变化



数据描述：由图可得，1 点到 8 点间用户的各行为都不怎么活跃，在午间 12 点和晚上 20、21 点左右浏览、加购、收藏均出现了小高峰，说明此时用户的活跃度很高，但购买量在晚间并不高，说明用户一般在白天购买行为较多；且浏览、加购、收藏这三者的趋势基本相同，同时具有相似功能的加购和收藏这两种行为，明显用户更偏向使用加购，但购买的行为并没有太大波动，相比前三者更加平缓。综合以上，作为平台可以在 1 点到 8 点间减少宣传力度，而在晚上 20 点到 24 点间增大宣传力度，由此减少成本，提高宣传效率。

6.用户的不同商品类目偏好

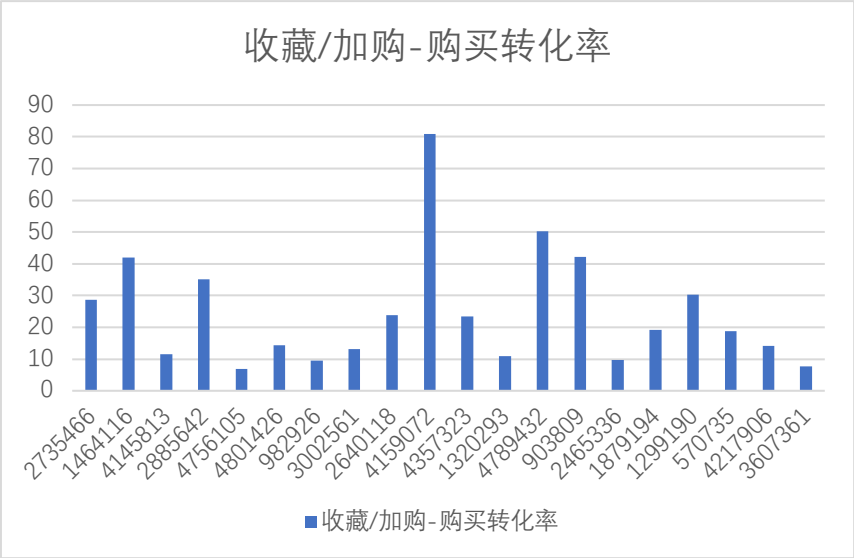
筛选出购买量前 20 名的商品类目用于分析。

```
# 用户对不同商品类目的偏好
# 筛选出排名前20的商品类目，同时展示其浏览，收藏加购数，再比较其浏览-购买转化率，收藏加购数转化率
with tmp as (select category as '商品类目',
    sum(case when behavior='pv' then 1 else 0 end) 浏览,
    sum(case when behavior='fav' then 1 else 0 end) 收藏,
    sum(case when behavior='cart' then 1 else 0 end) 购物车,
    sum(case when behavior='buy' then 1 else 0 end) 购买
from taobao
group by category
order by 购买 desc
) select 商品类目, round(购买/(收藏+购物车)*100,2) as '收藏加购-购买转化率', round(购买/(浏览)*100,2) as '浏览-购买转化率'
from tmp
limit 20
```

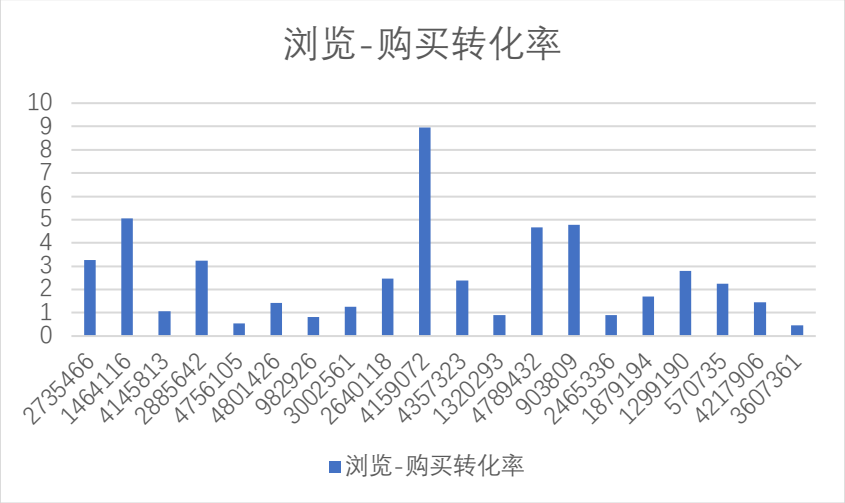
代码 9 计算排名前 20 商品类目的各行为的购买转化量

按 category 分组计算每类商品各用户行为的数量，并计算各类行为相对于购买行为的转化率。

图九 收藏/加购-购买转化率



图十 浏览-购买转化率



数据描述：其中可以看出，有些商品虽然购买量位于前 20，但其转化率却很低，也许是因为其有巨大的浏览数、曝光量。推测这种商品是比较廉价的、流量较高的网红商品，大家只是一时热度进去浏览，而购买欲望并不强烈。如果淘宝大量增加这样的商品的曝光度与宣传，就会容易让消费者模糊淘宝的定位人群，由此丧失一部分头部用户。同时，也推测商品热搜与热销并不匹配的情况可能是由于商品未精准匹配用户需求。

7.基于 RFM 模型划分用户

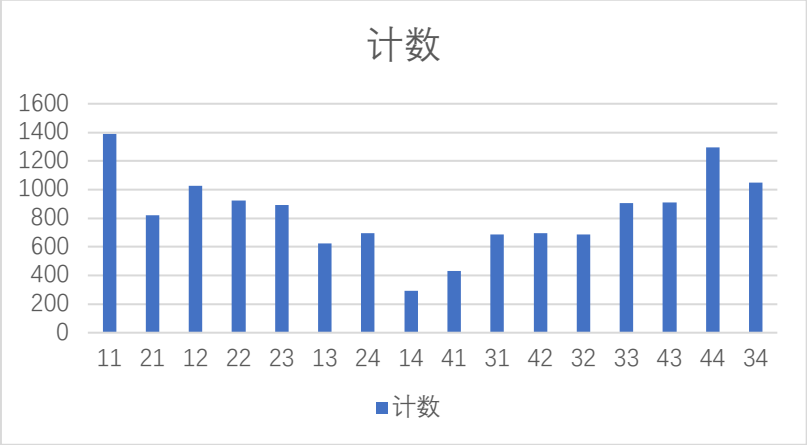
```
# 基于RFM模型找出有价值的用户
SET @total := (select count(DISTINCT user) from taobao where behavior = 'buy')

create table rr as
with tmp as (select USER , max(date) as rct, count(behavior) as frqc from taobao where behavior='buy'
group by USER)
select *, ROW_NUMBER() over(order by rct) as r from tmp

create table ff as
with tmp as (select USER , max(date) as rct, count(behavior) as frqc from taobao where behavior='buy'
group by USER)
select *, ROW_NUMBER() over(order by frqc) as f from tmp

select k.user_rf as user_rf, count(*) as cnt
from (select rr.user,concat(
case when rr.rr<@total/4 then '1'
when rr.rr>@total/4 and rr.rr<=@total/2 then '2'
when rr.rr>@total/2 and rr.rr<=@total/4*3 then '3'
when rr.rr>@total/4*3 and rr.rr<=@total then '4' end,
case when ff.ff<=@total/4 then '1'
when ff.ff>@total/4 and ff.ff<=@total/2 then '2'
when ff.ff>@total/2 and ff.ff<=@total/4*3 then '3'
when ff.ff>@total/4*3 and ff.ff<=@total then '4' end) as user_RF
from rr join ff
on
rr.user = ff.user) as k
group by k.user_rf
```

代码 10 RFM 模型划分价值用户

首先，创建一个变量 total，令其等于所有购买的用户数。再 max(date)计算得到每位用户最近一次有购买行为的时间，然后按用户划分去计算每位用户有购买行为的次数，记为 frqc。然后分别按 recency 和 frequency 增序排序，得到 rr 表和 ff 表。

接着进行打分操作，将 rr 表和 ff 表中的 recency 和 frequency 序号与 total 相比较，其中，用户被平等划分成四个等级。recency 或 frequency 的序号越大的，用户越具有价值，依次赋分为 1、2、3、4。赋分后，将两个数字直接进行连接，得到用户价值最终评级。

图十一 RFM 模型各类用户计数图



数据描述：从该图中并没有看出很明显淘宝的用户划分的规律，但总体来讲，淘宝大部分的用户还是比较成熟，消费频率和最近购买时间都还较有价值。

进一步的分析可以采用更多维度的数据，比如消费总金额，对于各价值等级的用户进行用户画像，以便更好的定位投送特定的商品，减少成本的同时提高宣传的精准度。