淘宝 APP 用户行为数据分析

一. 了解数据

- **数据说明：**数据集包含了 2014 年 11 月 18 日至 2014 年 12 月 18 日之间，一个月内的淘宝 APP 移动端用户行为数据。
- **数据来源：**阿里云天池 https://link.zhihu.com/?target=https%3A//tianchi.aliyun.com/dataset/dataDetail%3FdataId%3D649%26userId%3D1
- **数据字段含义：**

字段	字段说明	提取说明
user_id	用户身份	抽样&字段脱敏
item_id	商品ID	字段脱敏
behavior_type	用户行为类型	包含浏览、收藏、加购物车、购买四种行为，分别用数字1、2、3、4表示
user_geohash	用户的地理位置，可以为空	通过经纬度算法生成
item_category	商品类别	字段脱敏
time	用户行为发生时间	精确到小时

二. 提出问题

1、整体用户的购物情况

pv（总访问量）、日均访问量、uv（用户总数）、有购买行为的用户数量、用户的购物情况、复购率分别是多少?

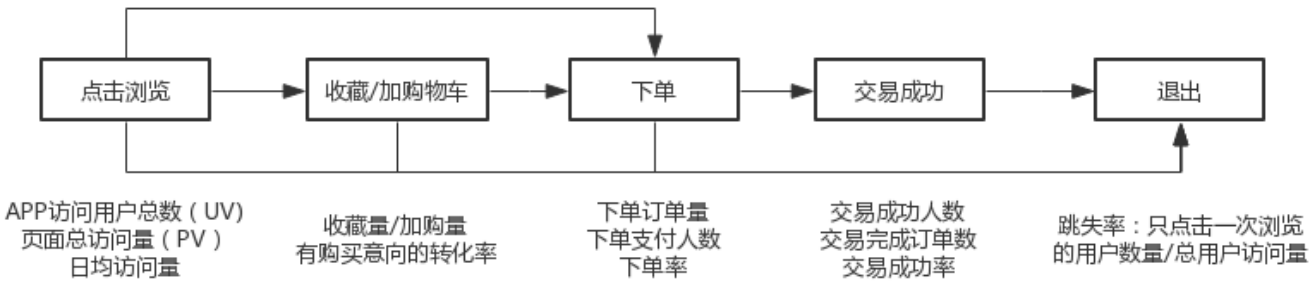
2、用户行为转化漏斗

点击 → 加购物车/收藏 → 购买各环节转化率如何？购物车遗弃率是多少，如何提高?

3、购买率高和购买率为 0 的人群有什么特征?

4、基于时间维度了解用户的行为习惯

5、基于 RFM 模型的用户分析



三. 数据清洗

- 使用的工具： WorkBench

1、删除重复值

数据集中用户的行为时间只精确到小时，所以会存在部分用户在某一小时内重复浏览、收藏或购买同一商品的行为，从而生成重复数据记录，对该数据去重。

创建数据库表是对 user_id、item_id、behavior_type 和 time 列定义主键约束，确保导入的数据无重复值。

```
# 导入非重复数据
Create Table data.tbuser(
    user_id Int(11) Not Null,
    item_id Int(11) Not Null,
    behavior_type Varchar(45) Not Null,
    user_geohash Varchar(45) Default Null,
    item_category Varchar(45) Default Null,
    time Varchar(45) Not Null,
    Primary Key (user_id, item_id, behavior_type, time))
Engine = Innodb
Default Charset = UTF8;
```

2、缺失值处理

user_geohash 列地理位置的数据大多是空值 NULL，且做过脱敏处理，难以补充和研究，将其去除不做和位置有关的分析。

其余数据中没有缺失值。

```
# 删除位置列
Alter Table data.tbuser Drop Column user_geohash;
```

3、一致化处理

time 字段的时间包含（年-月-日）和小时，为方便分析，将该字段拆分成两列—日期列和时间列。

```
# 在time前面插入一列date
Alter Table data.tbuser Add date Varchar(20) Not Null Before time;

# 截取日期
Update data.tbuser Set date = Substring_Index(time,' ',1);

# 在date后面插入一列hours
Alter Table data.tbuser Add hours Varchar(20) Not Null After date;

# 截取时间
Update data.tbuser Set house = Substring(time from 12 for 2);
```

```
# 修改日期列的数据类型
Alter Table data.tbuser Modify date Date;
```

behavior_type 列中分别用 1，2，3，4 表示点击、收藏、加购物车、购买四种行为类型，为方便查看，将 1，2，3，4 替换为 ‘pv’，’ fav ‘，’ cart'， ‘buy' 。

```
# 将 behavior_type 列数据替换
Update data.tbuser Set behavior_type = Replace(behavior_type,'1','pv');

Update data.tbuser Set behavior_type = Replace(behavior_type,'2','fav');

Update data.tbuser Set behavior_type = Replace(behavior_type,'3','cart');

Update data.tbuser Set behavior_type = Replace(behavior_type,'4','buy');
```

若更改列值时报错，可尝试如下语句：

```
SET SQL_SAFE_UPDATES = 0   # 关闭 safe_updates 模式，防止报错，开启为 1
```

4、异常值处理

检查日期是否在规定范围内：2014 年 11 月 18 日至 2014 年 12 月 18 日。

```
# 查询表结构
Desc DATA.tbuser;
```

Field	Type	Null	Key	Default	Extra
user_id	int(11)	NO	PRI	NULL	
item_id	int(11)	NO	PRI	NULL	
behavior_type	varchar(45)	NO	PRI	NULL	
item_category	varchar(45)	YES		NULL	
date	date	YES		NULL	
hours	varchar(20)	NO		NULL	
time	varchar(45)	NO	PRI	NULL	

```
# 查询日期范围
Select Max(date), Min(date) From project.tbuser;
```

max(date)	min(date)
2014-12-18	2014-11-18

数据符合规定范围，无需删除数据

- 完成数据清洗后的数据：

user_id	item_id	behavior_type	item_category	date	hours	time
4913	315532	pv	3431	2014-12-16	12	2014-12-16 12
4913	876969	pv	4582	2014-11-27	14	2014-11-27 14
4913	876969	pv	4582	2014-11-27	18	2014-11-27 18
4913	876969	pv	4582	2014-11-27	19	2014-11-27 19
4913	876969	pv	4582	2014-11-30	23	2014-11-30 23
4913	2741340	pv	1308	2014-11-30	21	2014-11-30 21

四. 构建模型和分析问题

1、 总体用户购物情况

1) pv（总访问量）

```
Select Count(behavior_type) as 总访问量 From data.tbuser
     Where behavior_type = 'pv';
```

总访问量
5535879

2) 日均访问量

```
Select date,Count(behavior_type) as 日均访问量 From data.tbuser
     Where behavior_type = 'pv'
     Group by date
     Order by date;
```

date	日均访问量
2014-11-18	166575
2014-11-19	162025
2014-11-20	159409
2014-11-21	151353
2014-11-22	163625
2014-11-23	173846
2014-11-24	171186

3) uv（用户总数）

```
Select Count(distinct(user_id)) as 用户总数 From data.tbuser;
```

用户总数
10000

4) 有购买行为的用户数量

```
Select Count(distinct(user_id)) as 购买用户数 From data.tbuser
     Where behavior_type = 'buy';
```

购买用户数
8886

5) 用户的购物情况

```
Create View user_behavior as
Select user_id,Count(behavior_type) as 用户行为总量,
     Sum(Case When behavior_type = 'pv' Then 1 Else 0 End) as 点击数,
     Sum(Case When behavior_type = 'fav' Then 1 Else 0 End) as 收藏数,
     Sum(Case When behavior_type = 'cart' Then 1 Else 0 End) as 加购数,
     Sum(Case When behavior_type = 'buy' Then 1 Else 0 End) as 购买数
From data.tbuser
     Group by user_id
     Order by Count(behavior_type) Desc;

Select * From user_behavior;
```

user_id	用户行为总里	点击数	收藏数	加购数	购买数
36233277	15936	12666	2909	318	43
73196588	15627	15627	0	0	0
65645933	12427	9823	2563	27	14
59511789	11196	9642	1262	262	30
7234861	9053	7276	1163	566	48
83813302	8199	7826	277	81	15
130270245	7920	6994	726	182	18
52772551	7326	7081	188	39	18

6) 复购率：产生两次或两次以上购买的用户占购买用户的比例

```
Select Sum(复购用户) as 复购数,Count(复购用户) as 总购买数,
     Concat(Round(Sum(复购用户)/Count(复购用户)*100,2),"%") as 复购率
From (Select user_id,Count(behavior_type),
           If(Count(behavior_type)>1,1,0) as 复购用户
      From data.tbuser
           Where behavior_type = 'buy'
           Group by user_id) as f;
```

复购数	总购买数	复购率
8117	8886	91.35%

2、 用户行为转化漏斗

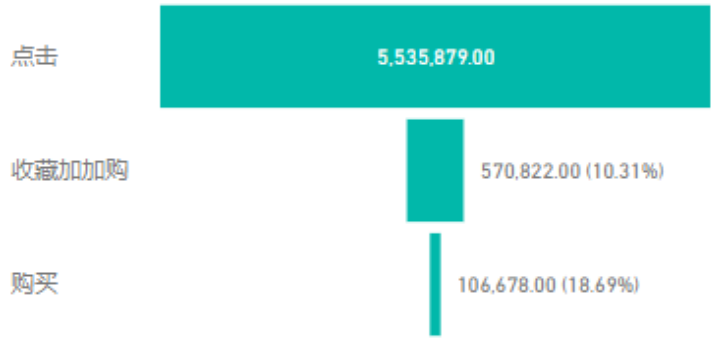
在购物环节中收藏和加入购物车两个环节没有先后之分，所以将这两个环节可以放在一起作为购物环节的一步。最终得到用户购物行为各环节转化率，如下：

```
# 用户购买行为统计
Select Sum(点击数) as 总点击数,Sum(收藏数) as 总收藏数,
     Sum(加购数) as 总加购数,Sum(购买数) as 总购买数
From user_behavior;
```

总点击数	总收藏数	总加购数	总购买数
5535879	239472	331350	106678

```
# 用户购买行为转化率
Select
     Concat(Round((Sum(收藏数)+Sum(加购数))/Sum(点击数)*100,2),'%') as
pv_to_favcart,
     Concat(Round(Sum(购买数)/(Sum(收藏数)+Sum(加购数))*100,2),'%') as
fav_to_buy,
     Concat(Round(Sum(购买数)/Sum(点击数)*100,2),'%') as pv_to_buy
From user_behavior;
```

pv_to_favcart	fav_to_buy	pv_to_buy
10.31%	18.69%	1.93%



不同的行业转化率有所差异，据2012 年的一项研究表明，在整个互联网范围内，点击到购买的平均转化率为2.13%（数据来源于《精益数据分析》），图中所示购买行为的转化率（点击→购买）为 1.93%，低于行业平均值，表明淘宝移动端用户行为的转化率依然有增长空间。

3、 购买率高和购买率为低的人群有什么特征

- 购买率高用户特征：

```
Select user_id,点击数,收藏数,加购数,购买数,
    Concat(Round(购买数/点击数*100,2),'%') as 购买率
From user_behavior
    Order by 购买率 Desc;
```

user_id	点击数	收藏数	加购数	购买数	购买率
56970308	31	0	30	30	96.77%
42281108	26	0	25	25	96.15%
117681959	342	11	35	34	9.94%
29760352	292	9	56	29	9.93%
123553065	121	2	2	12	9.92%
65982317	1292	2	260	128	9.91%
31185970	91	0	13	9	9.89%
58784511	162	2	15	16	9.88%
98224094	71	0	23	7	9.86%
112539954	305	2	47	30	9.84%

```
Select user_id,点击数,收藏数,加购数,购买数,
    Concat(Round(购买数/点击数*100,2),'%') as 购买率
From user_behavior
    Order by 购买数 Desc;
```

user_id	点击数	收藏数	加购数	购买数	购买率
122338823	4262	6	903	745	17.48%
123842164	4944	164	536	383	7.75%
51492142	4136	206	592	332	8.03%
56560718	2522	311	362	238	9.44%
33448326	1402	37	266	187	13.34%
35306096	1761	96	549	163	9.26%
93675262	1878	133	131	143	7.61%
100816273	1795	15	449	140	7.80%
60674956	638	70	150	136	21.32%
4120403	1793	70	196	131	7.31%

由以上结果可以看出，购买率高的用户点击率反而不是最多的，这些用户收藏数和加购物车的次数也很少，一般不点击超过 5 次就直接购买，由此可以推断出这些用户为理智型消费者，有明确的购物目标，属于缺啥买啥型，很少会被店家广告或促销吸引。

- 购买率为低用户特征：

```
Select user_id,点击数,收藏数,加购数,购买数,
    Concat(Round(购买数/点击数*100,2),'%') as 购买率
```

```
From user_behavior
    Order by 购买率;
```

user_id	点击数	收藏数	加购数	购买数	购买率
45021776	22	0	2	0	0.00%
37409400	960	115	44	0	0.00%
29533214	19	4	0	0	0.00%
29554512	22	0	1	0	0.00%
27858325	19	4	0	0	0.00%
53281938	72	9	4	0	0.00%
68790834	22	0	0	0	0.00%
27914256	22	0	0	0	0.00%
111244765	17	0	4	0	0.00%
123746892	16	0	5	0	0.00%
61775192	18	0	3	0	0.00%
29025946	83	0	0	0	0.00%

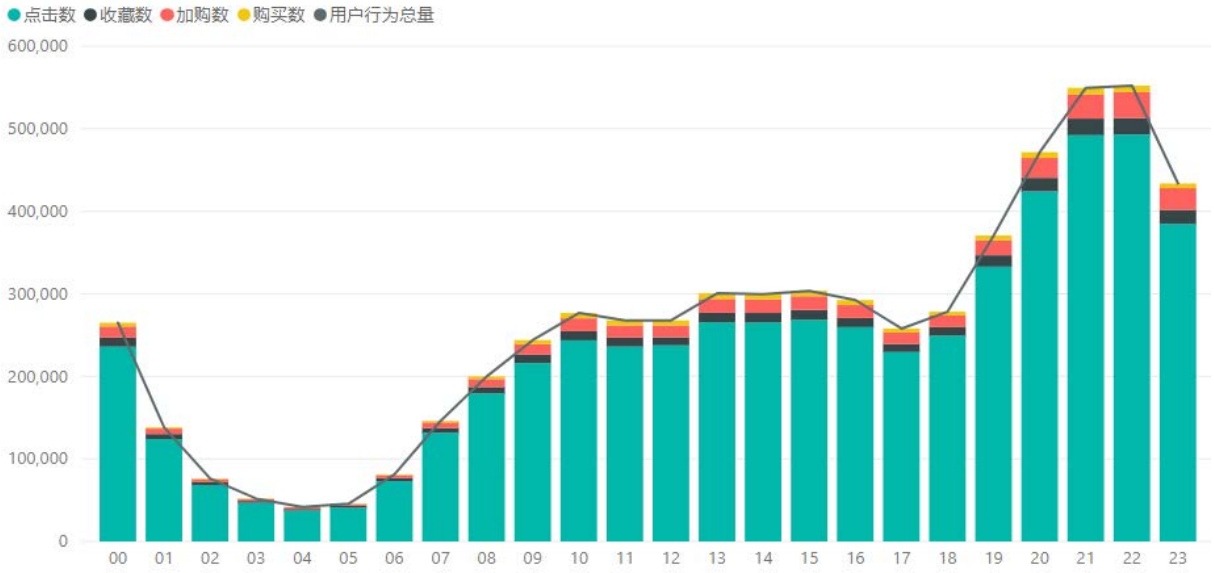
由以上结果可以看出，购买率为低用户分为两类，一类是点击次数少的，一方面的原因是这类用户可能是不太会购物或者不喜欢上网的用户，可以加以引导，另一方面是从商品的角度考虑，是否商品定价过高或设计不合理；第二类用户是点击率高、收藏或加购物车也多的用户，此类用户可能正为商家的促销活动做准备，下单欲望较少且自制力较强，思虑多或者不会支付，购物难度较大。

4、 基于时间维度了解用户的行为习惯

1) 一天中用户的活跃时段分布

```
Select hours,Count(behavior_type) as 用户行为总量,
    Sum(Case When behavior_type = 'pv' Then 1 Else 0 End) as 点击数,
    Sum(Case When behavior_type = 'fav' Then 1 Else 0 End) as 收藏数,
    Sum(Case When behavior_type = 'cart' Then 1 Else 0 End) as 加购数,
    Sum(Case When behavior_type = 'buy' Then 1 Else 0 End) as 购买数
From data.tbuser
    Group by hours
    Order by hours;
```

hours	用户行为总量	点击数	收藏数	加购数	购买数
00	264993	235979	10957	13652	4405
01	138190	123931	6231	6537	1491
02	76170	68462	3279	3703	726
03	51872	46794	2250	2378	450
04	41865	37445	1962	2091	367
05	45623	41025	2029	2147	422
06	81224	73035	3590	3658	941
07	146054	131642	5813	6817	1782
08	199962	179365	7730	9644	3223
09	244014	216149	10320	12528	5017
10	276815	243655	11072	15616	6472



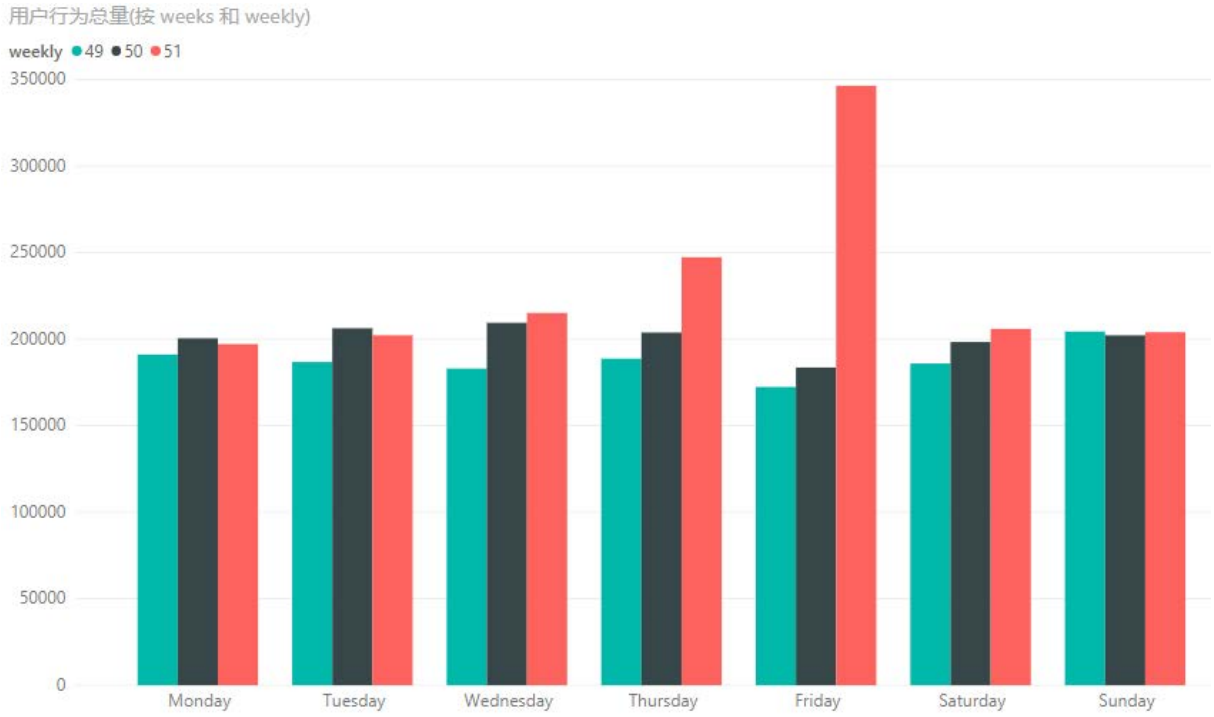
可以看出，每日 0 点到 5 点用户活跃度快速降低，降到一天中的活跃量最低值，6 点到 10 点用户活跃度快速上升，10 点到 18 点用户活跃度较平稳，17 点到 23 点用户活跃度快速上升，达到一天中的最高值。

2) 一周中用户活跃时段分布

由于第一周和第五周的数据不全，因此这两周的数据不考虑到此次分析中。

```
Select weekly,weeks,Count(behavior_type) as 用户行为总量,
     Sum(Case When behavior_type = 'pv' Then 1 Else 0 End) as 点击数,
     Sum(Case When behavior_type = 'fav' Then 1 Else 0 End) as 收藏数,
     Sum(Case When behavior_type = 'cart' Then 1 Else 0 End) as 加购数,
     Sum(Case When behavior_type = 'buy' Then 1 Else 0 End) as 购买数
From (Select date,Week(Date_Add(Min(date),Interval 1 week),1) as weekly,
           Date_Format(date,'%W') as weeks
      From data.tbuser
           Where date Between '2014-11-24' and '2014-12-14'
           Group by date) as w, data.tbuser
     Where w.date = data.tbuser.date
     Group by weekly,weeks
     Order by weekly,Field(weeks,'Monday','Tuesday','Wednesday','Thursday',
     'Friday','Saturday','Sunday'); # 自定义排序
```

weekly	weeks	用户行为总量	点击数	收藏数	加购数	购买数
49	Monday	191139	171186	7149	9750	3054
49	Tuesday	186803	167036	7280	9401	3086
49	Wednesday	183017	163529	7277	9084	3127
49	Thursday	188747	168221	7660	9611	3255
49	Friday	172392	154289	6427	8800	2876
49	Saturday	185940	166197	7130	9754	2859
49	Sunday	204398	183004	7737	10398	3259
50	Monday	200520	178919	7269	10905	3427
50	Tuesday	206313	183953	7949	11159	3252
50	Wednesday	209477	186225	8436	11375	3441



可以看出，各周用户活跃度较稳定，每周五会有小幅降低，但会在周末慢慢回升到原活跃度范围。其中第三周周五（2014-12-12）用户活跃度突增，主要是受到了双十二电商大促销活动的影响。

5、 基于 RFM 模型找出有价值的用户

RFM 模型是衡量客户价值和客户创利能力的重要工具和手段，其中由 3 个要素构成了数据分析最好的指标，分别是：

R-Recency（最近一次购买时间）

F-Frequency（消费频率）

M-Money（消费金额）

由于数据源没有相关的金额数据，暂且通过 R 和 F 的数据对客户价值进行打分。

1） 计算 R-Recency

由于数据集包含的时间是从 2014 年 11 月 18 日至 2014 年 12 月 18 日，这里选取 2014 年 12 月 19 日作为计算日期，统计客户最近发生购买行为的日期距离 2014 年 12 月 19 日间隔几天，再对间隔时间进行排名，间隔天数越少，客户价值越大，排名越靠前。

```
Select a.*,(@rank :=@rank+1) as Recency_rank
From (Select user_id,Max(date),Datediff('2014-12-19',Max(date)) as Recency
     From data.tbuser
          Where behavior_type='buy'
          Group by user_id
          Order by Recency) a,(Select @rank :=0) b;
```

user_id	Recency_date	Recency	Recency_rank
62781637	2014-12-18	1	1
63052028	2014-12-18	1	2
63339028	2014-12-18	1	3
63702010	2014-12-18	1	4
128289511	2014-12-18	1	5
65762927	2014-12-18	1	6
129706841	2014-12-18	1	7
130497558	2014-12-18	1	8
867381	2014-12-18	1	9
1091123	2014-12-18	1	10

2) 计算 F-Frequency

先统计每位用户的购买频率，再对购买频率进行排名，频率越大，客户价值越大，排名越靠前。

```
Select a.*,(@rank :=@rank+1) as Frequency_rank
From (Select user_id,count(user_id) as Frequency From data.tbuser
            Where behavior_type='buy'
            Group by user_id
            Order by Frequency Desc) a,(Select @rank :=0) b;
```

user_id	Frequency	Frequency_rank
122338823	745	1
123842164	383	2
51492142	332	3
56560718	238	4
33448326	187	5
35306096	163	6
93675262	143	7
100816273	140	8
60674956	136	9
4120403	131	10

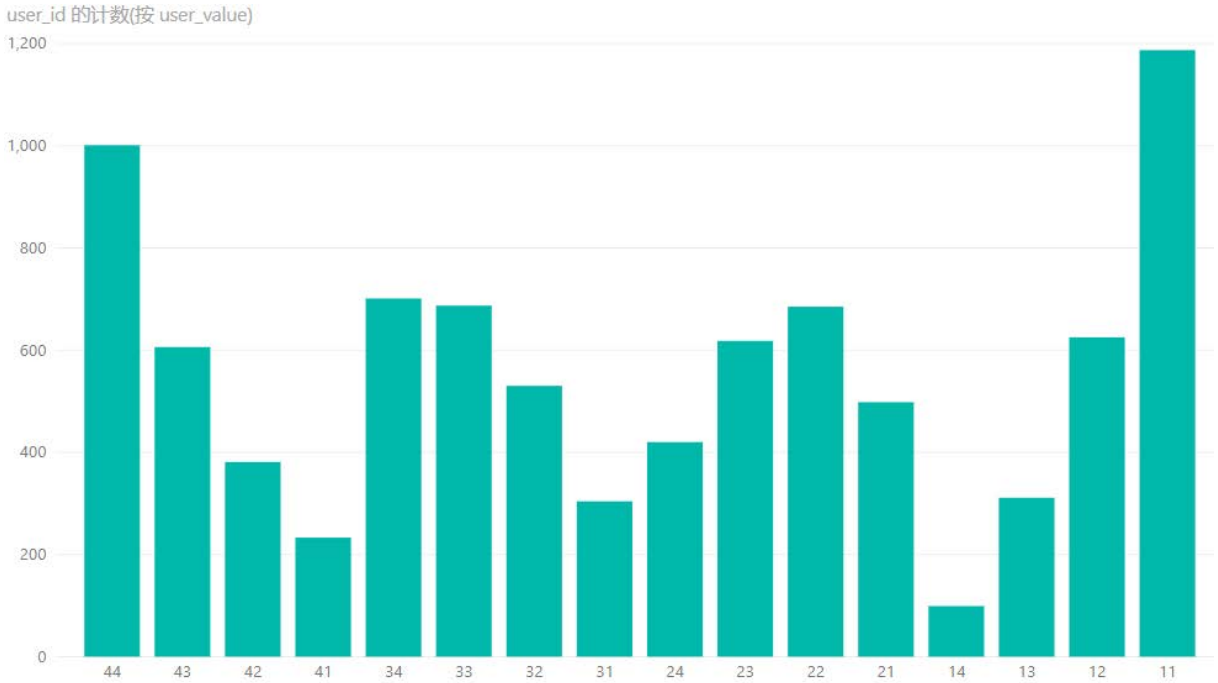
3) 对用户进行评分

对 8886 名有购买行为的用户按照排名进行分组，共划分为四组，对排在前四分之一的用户打 4 分，排在前四分之一到四分之二（即二分之一）的用户打 3 分，排在前四分之二到前四分之三的用户打 2 分，剩余的用户打 1 分，按照这个规则分别对用户时间间隔排名和购买频率排名打分，最后把两个分数合并在一起作为该名用户的最终评分。

```
Select r.user_id,r.Recency_date,r.Recency,r.Recency_rank,f.Frequency,
     f.Frequency_rank,
     # 对客户购买行为的日期排名和频率排名进行打分
     Concat(Case When Recency_rank<=8886/4 Then '4'
            When Recency_rank<=8886/2 and Recency_rank>8886/4 Then '3'
            When Recency_rank<=8886/4*3 and Recency_rank>8886/2 Then '2'
            Else '1' End,
            Case when Frequency_rank<=8886/4 Then '4'
            When Frequency_rank<=8886/2 and Frequency_rank>8886/4 Then '3'
            When Frequency_rank<=8886/4*3 and Frequency_rank>8886/2 Then '2'
            Else '1' End) as user_value
From
```

```sql
# 对每位用户最近发生购买行为的间隔时间进行排名（间隔天数越少，客户价值越大）
(Select a1.*, (@rank1 :=@rank1 + 1) as Recency_rank From
 # 统计客户最近发生购买行为的日期距离'2014-12-19'间隔几天
      (Select user_id,Max(date) as Recency_date,
           Datediff('2014-12-19', Max(date)) as Recency
       From data.tbuser
            Where behavior_type = 'buy'
            Group by user_id
            Order by Recency) a1, (Select @rank1 :=0) b1) r,
 # 对每位用户的购买频率进行排名（频率越大，客户价值越大）
 (Select a2.*, (@rank2 :=@rank2 + 1) as Frequency_rank From
 # 统计每位用户的购买频率
      (Select user_id, Count(behavior_type) as Frequency
       From data.tbuser
            Where behavior_type = 'buy'
            Group by user_id
            Order by Frequency Desc) a2, (Select @rank2 :=0) b2) f
Where r.user_id = f.user_id
Group by r.user_id;
```

user_id	Recency_date	Recency	Recency_rank	Frequency	Frequency_rank	user_value
4913	2014-12-16	3	3086	5	6043	32
6118	2014-12-17	2	2066	1	8825	41
7528	2014-12-13	6	5160	6	5572	22
7591	2014-12-13	6	5198	19	1552	24
12645	2014-12-14	5	4715	8	4179	23
54056	2014-12-07	12	7445	2	7522	11
63348	2014-12-11	8	6777	1	8282	11
79824	2014-12-16	3	3480	13	2580	33
88930	2014-12-17	2	2566	19	1593	34
100539	2014-12-16	3	3583	18	1727	34



给用户的购买时间间隔和和购买频率打分有助于了解每位顾客的特性，从而实现差异化营销。比如对 user_value = 44 的重要用户需要重点关注；对 user_value = 41 这类忠诚度高而购买能力不足的客户，可以给与适当折扣或采用捆绑销售来增加用户的购买频率。对 user_value = 14 这类忠诚度不高而购买能力强的客户，需要关注他们的购物习性做精准化营销。

另外，还可通过每个月用户的评分变化，推测客户消费的异动状况，对即将流失的客户，通过电话问候、赠送礼品、加大折扣力度等有效的方式进行挽回。

五. 结论

1、总体转化率只有 1.93%，用户点击后收藏和加购物车的转化率在 10.31%，需要提高用户的购买意愿，可采用活动促销、精准营销等方式。

2、购买率高且点击量少的用户属于理智型购物者，有明确购物目标，受促销和广告影响少；而购买率低的用户可以认为是等待型或克制型用户群体，下单欲望较少且自制力较强，购物难度较大。

3、大部分用户的主要活跃时间在 10 点到 23 点，在 19 点到 23 点达到一天的顶峰。每周五的活跃度有所下降，但周末开始回升。可以根据用户的活跃时间段精准推送商家的折扣优惠或促销活动，提高购买率。

4、通过 R 和 F 的数据给用户行为打分，对每位用户进行精准化营销，还可以通过对 R 和 F 的数据监测，推测客户消费的异动状况，挽回流失客户。